Usytuowanie na poziomie samorządów lokalnych instrumentów wsparcia dla MŚP, działających w oparciu o model wielopoziomowego zarządzania regionem REGIOGMINA

Model prognozy wskaźnika innowacyjności gminy

dr Karol Przanowski
Szkoła Główna Handlowa - SGH

Projekt finansowany z Narodowego Centrum Badań i Rozwoju w ramach w ramach Programu Strategicznego „Społeczny i gospodarczy rozwój Polski w warunkach globalizujących się rynków" GOSPOSTRATEG

Warszawa, lipiec 2020

© Copyright by dr Karol Przanowski
Szkoła Główna Handlowa - SGH
Warszawa, lipiec 2020

Spis treści

Cel budowy modelu **4**

1 Struktury danych **5**
1.1 Definicja zdarzenia modelowego 5
1.2 Konstrukcja zmiennych modelowych 6
1.3 Metodyka budowy modelu 6

2 Podstawowe informacje o modelu **9**
2.1 Karta skoringowa 9
2.2 Ranking gmin i metody polepszenia ich innowacyjności 10
2.3 Zbiory danych wykorzystane w estymacji 11
2.4 Opisy zmiennych 12
2.5 Statystyki zmiennych 19
2.6 Statystyki modelu 22
2.7 Kalibracja do wartości prognozy wskaźnika innowacyjności 24
2.8 Analiza jakości zmiennych 24

Bibliografia **30**

Spis rysunków **32**

Spis tabel **33**

Cel budowy modelu

W ramach Projektu „Gospostrateg” zrodziła się idea budowy „Obserwatorium Gospodarczego”, które powinno pomagać gminom i ich władzom samorządowym wspierać wszelkie inicjatywy zmierzające do rozwoju przedsiębiorczości i innowacyjności w województwie Kujawsko-Pomorskim.

Zadanie pierwsze projektu pozwoliło zebrać dostępne dane obiektywne o przedsiębiorstwach ze źródeł takich jak: REGON, GUS – Bank Danych Lokalnych, ZUS i Urząd Skarbowy. W ramach zdania drugiego w projekcie przeprowadzono ankietę wśród przedsiębiorców i zebrano dane subiektywne oparte na deklaracji ankietowanych.

Posiadanie obu rodzaju danych: obiektywnych i subiektywnych pozwoliło na wykonanie oryginalnego eksperymentu połączenia ich razem i zbudowanie modelu statystycznego, który na podstawie danych obiektywnych, możliwych do cyklicznego pobierania, pozwala prognozować wskaźnik innowacyjności oparty na jednorazowym przeprowadzeniu ankiety i subiektywnych odpowiedziach przedsiębiorców.

Tak zbudowany model pozwala prognozować i mierzyć innowacyjność gmin co roku, bez konieczności przeprowadzania kolejnych ankiet na dużą skalę. Pomaga też definiować kluczowe wskaźniki efektywności działań gmin i samorządów, co w konsekwencji może prowadzić do definiowania właściwych strategii, celów (nawet ilościowych) i skierować uwagę gmin na mierzalne efekty i działania.

Model ten doskonale zatem spełnia kryteria do stosowania go właśnie w „Obserwatorium Gospodarczym”, jako narzędzia do pomiaru przedsiębiorczości i innowacyjności przedsiębiorstw w gminach.

1. Struktury danych

1.1. Definicja zdarzenia modelowego

Definicja zdarzenia modelowego jest jednym z kluczowych pojęć przy budowie modelu predykcyjnego.

Do identyfikacji zdarzenia modelowego posłużyła ankieta przeprowadzona w zadaniu drugim projektu Gospostrateg, która dotyczyła badania empirycznego przeprowadzonego w ramach zadania drugiego fazy badawczej. Zostało ono zrealizowane przez zespół badawczy składający się z pracowników Wydziału Nauk Ekonomicznych i Zarządzania Uniwersytetu Mikołaja Kopernika w Toruniu, kierowanego przez prof. dr hab. Macieja Zastempowskiego. Celem zadania było sporządzenie statystycznych i ekonometrycznych analiz dla MŚP województwa kujawsko-pomorskiego. Zespół badawczy składał się z następujących osób: dr I. Escher, dr hab. W. Glabiszewski, prof. UMK, prof. A. Glińska-Neweś, dr D. Grego-Planer, dr K. Liczmańska-Kopcewicz oraz dr hab. E. Wędrowska, prof. UMK.

Każde przedsiębiorstwo odpowiedziało na następujące pytanie - B5: Czy wprowadziliście nowe lub istotnie ulepszone metody wytwarzania produktów lub usług?

Wyróżniamy następujące statusy funkcji celu do modelowania:

- **Inno** – przedsiębiorstwo innowacyjne, które odpowiedziało twierdząco na pytanie z ankiety.
- **Nieinno** – przedsiębiorstwo nieinnowacyjne, które odpowiedziało przecząco na pytanie z ankiety.
- **Nieok** – przedsiębiorstwo nieokreślone, w tym wypadku nie wykorzystano tego oznaczenia.

Statystykę obserwowaną, która opisuje udział innowacyjnych firm w danej grupie (gminie) oznaczamy jako Inn. Prognozowaną jej wartość oznaczamy jako P. inno.

1.2. Konstrukcja zmiennych modelowych

Dane do konstrukcji zmiennych opisujących przedsiębiorstwa pochodzą z zagregowanych wskaźników na poziomie gminy z następujących źródeł:

- **REGON** – liczebności przedsiębiorstw w latach 2017-2018.
- **GUS Bank Danych Lokalntch** – bezrobocie, zobowiązania, liczba ludności w latach 2017-2018.
- **ZUS** – liczby płatników i sumy składek w latach 2017-2018 w rozbiciu na wielkość przedsiębiorstw, PKD i grupy wiekowe.
- **Urząd Skarbowy** – pozycje zeznań PIT36, CIT8, PIT28A oraz VAT w latach 2017-2018 w rozbiciu na wielkość przedsiębiorstw i grupy wiekowe.

Zdarzenie modelowe związane jest z ankietą przeprowadzoną w pod koniec roku 2019. Pozwala to na przeprowadzenie klasycznej analizy przyczynowo skutkowej polegającej na poszukiwaniu takich cech gmin z lat 2017-2018, że przyczyniły się do zwiększenia lub zmniejszenia udziału innowacyjnych firm w roku 2019.

Finalnie zostało zbudowanych 396 zmiennych, które głównie są udziałami lub stosunkami szczegółowych wielkości do ogólnych na gminę.

Temat przygotowania danych i rozbudowy ich zakresu jest bardzo pożądany. Obecna wersja modelu oparta jest na pierwszej iteracji transferu danych ze źródeł do finalnego przeliczenia modelu. Zebrane doświadczenie może pomóc w kolejnych etapach doskonalenia modelu, gdyż im więcej danych, tym model będzie lepszy i uda się jeszcze lepiej identyfikować czynniki wpływające na innowacyjność przedsiębiorstw.

1.3. Metodyka budowy modelu

Metodyka budowy modelu jest zgodna z podstawową teorią opisaną w dokumentach Basel II, takimi jak (BIS-WP14, 2005; BIS-BASEL, 2005). Budowa modeli jest także oparta na podstawowych metodach

opisanych w literaturze przedmiotu odwołujacej się głównie do książek i artykułów (Janc i Kraska, 2001; Siddiqi, 2005; Thonabauer i Nosslinger, 2004; Lessmanna *et al.*, 2013; Thomas *et al.*, 2002; Verstraeten i den Poel, 2005; Anderson, 2007). Materiał rozszerzający, uzupełniający i porządkujący całą metodykę opracowany jest w (Przanowski, 2014, 2015).

Proces budowy modelu przebiega przez następujące etapy:

- Próby losowe. Stworzone zastały dwa zbiory: treningowy i walidacyjny metodą losowania prostego bez zwracania w proporcji $60\%/40\%$.

- Tworzenie kategorii – kategoryzacja zmiennych lub grupowanie (ang. binning). Na podstawie miary entropii każda zmienna ciągła jest kategoryzowana do zmiennej porządkowej. Jest to typowa metoda stosowana przy drzewach decyzyjnych. Zmienne nominalne lub porządkowe także się grupuje, tworząc finalnie liczniejsze kategorie, bardziej reprezentatywne w populacji.

- Wstępna selekcja zmiennych (preselekcja) – odrzucenie zmiennych o niskiej predykcyjności. W tym etapie na podstawie prostych jednowymiarowych kryteriów odrzuca się zmienne nienadające się do finalnego modelu, które mają albo znacząco słabą wartość predykcji, albo są bardzo niestabilne. Po kategoryzacji wszystkie zmienne są transponowane do logitów na kategoriach.

- Wielowymiarowa selekcja zmiennych – generator modeli. Dla zmiennych ciągłych w procedurze Logistic istnieje metoda selekcji oparta na heurystyce podziału i ograniczeń (ang. branch and bound (Furnival i Wilson, 1974)).

- Ocena modeli. Nie ma jednego kryterium do oceny modelu. Stosuje się zatem kilka kryteriów głównie związanych z mocą predykcyjną (Gini), stabilnością (R. Gini, inaczej delta Gini – różnica względna predykcji pomiędzy zbiorami: treningowym i walidacyjnym), miary współliniowości: Max VIF –

maksymalny współczynnik inflacji wariancji[1] (ang. Variance Inflation Factor) (Koronacki i Mielniczuk, 2001), MAX Pearson – maksymalny współczynnik korelacji Pearsona na parach zmiennych i MAX Con Index – maksymalny indeks warunkujący (ang. Condition Index) (Welfe, 2003) oraz miara istotności.

- Wdrożenie modelu. Proponowane uruchomienie wyliczania wskaźników innowacyjności gmin jako cykliczny proces, który potem powinno się coraz bardziej ulepszać i jednocześnie na jego bazie wyciągać coraz to lepsze wnioski.

- Monitoring i testowanie. Każde nowe przeliczenie wskaźników powinno skłaniać do wykonania weryfikacji poprawności wyliczonych mierników. W tym celu powinno się przeprowadzać ankiety na mniejszą skalę, by nadal mieć zgodność prognozy z obserwowaną wartością. Dodatkowo możliwy jest ranking gmin i przeprowadzanie szczególnych analiz wśród gmin z największa innowacyjnością i najmniejszą.

[1] Czasem tłumaczony jako współczynnik podbicia wariancji.

2. Podstawowe informacje o modelu

2.1. Karta skoringowa

Pełna definicja karty skoringowej przedstawiona jest w tabeli 1. Domyślną kategorią jest zawsze kategoria z najmniejszą innowacyjnością, czyli w tym wypadku kategoria z najwyższą oceną cząstkową.

Tabela 1: Model prognozy innowacyjności

Gini w % na zbiorze		Lift1	Lift5	Lift10	Lift20
treningowym	walidacyjnym				
24,9	21,4	10,2	2,3	2,2	1,8

Karta skoringowa

Zmienna	Warunek	Ocena cząstkowa
ACT_MALE_SRPIT36_POZ109	11056.571429 $<$ ACT_MALE_SRPIT36_POZ109	59
	3352 $<$ ACT_MALE_SRPIT36_POZ109 $\leqslant$ 11056.571429	76
	ACT_MALE_SRPIT36_POZ109 $\leqslant$ 3352	85
ACT_MIKRO_PCTPIT28A_POZ10	ACT_MIKRO_PCTPIT28A_POZ10 $\leqslant$ 0.1983084509	59
	0.1983084509 $<$ ACT_MIKRO_PCTPIT28A_POZ10 $\leqslant$ 0.4762347181	67
	0.4762347181 $<$ ACT_MIKRO_PCTPIT28A_POZ10	69
ACT_MIKRO_PCTVAT_POZ19	ACT_MIKRO_PCTVAT_POZ19 $\leqslant$ 0.5465141947	59
	0.5465141947 $<$ ACT_MIKRO_PCTVAT_POZ19 $\leqslant$ 0.5572727784	63
	0.5572727784 $<$ ACT_MIKRO_PCTVAT_POZ19	68
ACT_PCT_PLAT_S	0.0668769716 $<$ ACT_PCT_PLAT_S	59
	0.0474683544 $<$ ACT_PCT_PLAT_S $\leqslant$ 0.0668769716	72
	ACT_PCT_PLAT_S $\leqslant$ 0.0474683544	74
ACT_SRED_PCTVAT_POZ27	0.0025255184 $<$ ACT_SRED_PCTVAT_POZ27	59
	0.0000389548 $<$ ACT_SRED_PCTVAT_POZ27 $\leqslant$ 0.0025255184	73
	ACT_SRED_PCTVAT_POZ27 $\leqslant$ 0.0000389548	79

Źródło: opracowanie własne.

2.2. Ranking gmin i metody polepszenia ich innowacyjności

Posiadając kartę skoringową można każdej gminie wyliczyć jej ocenę punktową za każdy rok rozliczeniowy (na bazie kolejnych zeznać PIT36, CIT8 i PIT28A). Dodatkowo dzięki kalibracji możliwe jest wyznaczenie nowego wskaźnika innowacyjności, który należy rozumieć jako prognozę procentowej liczby innowacyjnych przedsiębiorstw w gminie.

Bardzo ciekawym narzędziem jest także wyjaśnienie przyczyny większej lub mniejszej wartości tego wskaźnika dla danej gminy. Otóż metoda wyznaczania ocen cząstkowych jest tak dobrana, że gminie z największą innowacyjnością przyznaje najmniej punktów. Innymi słowy studiując postać karty skoringowej, dość szybko potrafimy wymienić własności gminy, która posiada największą innowacyjność i najmniejszą.

Jeśli zatem chcemy sprawić, by dana gmina miała większy wskaźnik innowacyjności to weryfikujemy wyznaczone dla niej oceny cząstkowe. Wszędzie tam, gdzie oceny te są duże, a można osiągnąć mniejsze, oznacza, że gmina musi zmienić wartość danej zmiennej, by przeskoczyć do innej kategorii z mniejszą liczbą punktów. Wystarczy zatem przeanalizować definicję kategorii, do której chcemy daną gminę zaklasyfikować i znaleźć krytyczną wartość zmiennej, którą gmina musi osiągnąć, by zmienić liczbę punktów.

Po identyfikacji kategorii i zmiennych należy przeprowadzić głębokie studia, by znaleźć istotne metody wpływające na zmianę wartości zmiennej, która jest najczęściej agregatem jakiejś wielkości liczonych na wszystkich przedsiębiorstwach w gminie. Ten temat wymaga dedykowanych warsztatów i spotkań interdyscyplinarnych pomagających rozwiązać problem. Pomimo wielu niewiadomych ta metoda pozwala jednak wyznaczyć konkretne działania i skierować uwagę z opinii eksperckich bez pokrycia na połączenie eksperckości z pomiarem ilościowym, który daje jednak pewną weryfikowalność.

2.3. Zbiory danych wykorzystane w estymacji

Model został zbudowany na podstawie zbioru danych ankietowanych przedsiębiorców w czasie czerwiec-wrzesień 2019. Natomiast dane o przedsiębiorcach wyliczono na podstawie danych z lat 2017 i 2018.

Liczby obserwacji w zbiorach treningowym i walidacyjnym przedstawione są w tabeli 2. Pełny zbiór danych całej dostępnej historii z liczebnościami po latach przedstawiony jest w tabeli 3.

Tabela 2: Liczebności zbiorów wykorzystanych w estymacji modelu

Zbiór	L. obserwacji	L. inno.	L. nieinno.	L. nieok.	P. inno. [%]	P. nieinno. [%]	P. nieok. [%]
Treningowy	3 170	508	2 662	0	16,0	84,0	0,0
Walidacyjny	1 666	294	1 372	0	17,6	82,4	0,0

Źródło: opracowanie własne.

Tabela 3: Liczebność zbioru wszystkich przedsiębiorstw

Rok	L. obserwacji	L. inno.	L. nieinno.	L. nieok.	P. inno. [%]	P. nieinno. [%]	P. nieok. [%]
2017	2 418	401	2 017	0	16,6	83,4	0,0
2018	2 418	401	2 017	0	16,6	83,4	0,0
Razem	4 836	802	4 034	0	16,6	83,4	0,0

Źródło: opracowanie własne.

2.4. Opisy zmiennych

Finalnie do modelu zostało wybranych tylko kilka najważniejszych zmiennych - wskaźników gmin. Wybór tych zmiennych został dokonany wyłącznie na podstawie kryteriów statystycznych, bez ingerencji ręcznej. Model ten należy traktować jako pierwszą wersję, dzięki czemu możliwe jest już prezentowanie jego efektów. W dalszej kolejności możliwe jest poświęcenie większej uwagi doborowi zmiennych, być może nawet poprzez cykl warsztatów w obecności wielu reprezentantów gmin, a także właścicieli danych (REGON, GUS, ZUS, Urząd Skarbowy), by wybrać zmienne bardziej intuicyjne i pomagające jeszcze lepiej wyznaczać kierunki rozwoju gmin.

Każdy raport zmiennej składa się z trzech elementów: raportu kategorii, wykresu stabilności udziałów kategorii i wykresu stabilności obserwowanego wskaźnika innowacyjności. Tak przedstawiona informacja jest bardzo przydatna biznesowo, gdyż pozwala zrozumieć reguły rządzące zachowaniem przedsiębiorstw, lub zauważyć własności gminy, które wpływają na zachowanie przedsiębiorstwa.

W raporcie kategorii najważniejszymi statystykami są: obserwowana innowacyjność (oznaczana jako Inn.) i udział. Pozwalają one szybko zauważyć najważniejsze prawidłowości. Dodatkowo prezentowana jest statystyka wartości informacyjnej (z ang. Information Value) oznaczana jako INV. Im większa jest jej wartość, tym zmienna i jej kategorie lepiej rozróżniają innowacyjne od nieinnowacyjnych przedsiębiorstw.

1. **ACT_MALE_SRPIT36_POZ109** – Średnie w gminie koszty uzyskania przychodu małżonka z pozarolniczej działalności gospodarczej w zeznaniu PIT36 (poz. 109, wariant 26) dla firm małych. Im większe są koszty uzyskania przychodu, tym większa jest innowacyjność w gminie, patrz rysunek 1

2. **ACT_MIKRO_PCTPIT28A_POZ10** – Stosunek sumarycznego przychodu w gminie z działalności prowadzonej na własne nazwisko opodatkowanego ryczałtem według stawki $8,5\%$ (poz. 10, wariant 18, PIT 28A) do przychodu według stawki $5,5\%$ (poz. 9, wariant 18, PIT 28A) dla firm mikro. Im więk-

szy jest ten stosunek, tym mniejsza innowacyjność w gminie, patrz rysunek 2

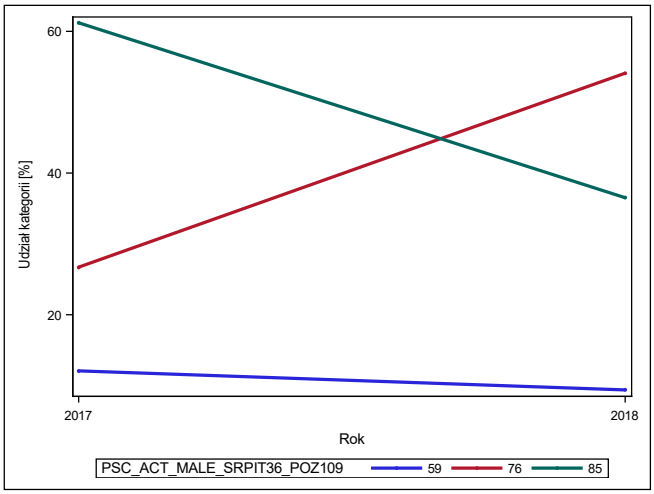
3. **ACT_MIKRO_PCTVAT_POZ19** – Stosunek sumy kwoty z poz. 19 (wariant 19) VAT, czyli dostawy towarów (sprzedaż) i świadczenie usług opodatkowanych stawką (22% albo 23%) do poz.40 - podstawy opodatkowania razem dla firm mikro. Im większy jest ten stosunek, tym mniejsza innowacyjność w gminie, patrz rysunek 3

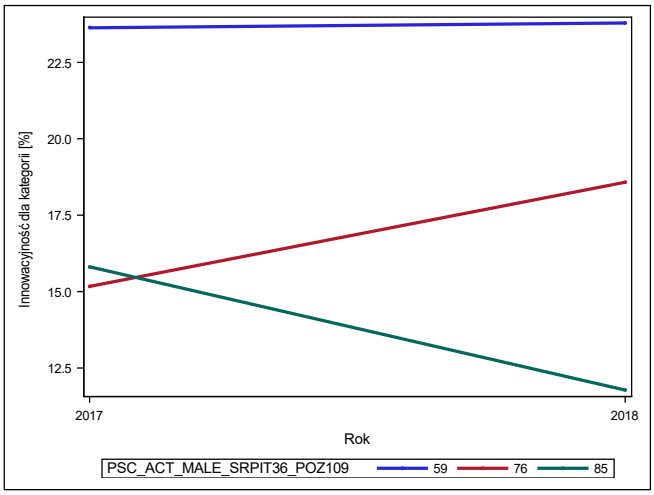
4. **ACT_PCT_PLAT_S** – Udział płatników ZUS w gminie z sekcji PKD S – Pozostała działalność usługowa w sumarycznej liczbie płatników ZUS w gminie. Im większy jest ten stosunek, tym większa jest innowacyjność w gminie, patrz rysunek 4

5. **ACT_SRED_PCTVAT_POZ27** – Stosunek sumy kwoty z poz. 27 (wariant 19) VAT, czyli importu usług z wyłączeniem usług nabywanych od podatników podatku od wartości dodanej, do których stosuje się art. 28b ustawy do poz.40 - podstawy opodatkowania razem dla firm średnich. Im większy jest ten stosunek, tym większa jest innowacyjność w gminie, patrz rysunek 5

Rysunek 1: Kategorie zmiennej ACT_MALE_SRPIT36_POZ109

Ocena	Warunek	Inn. [%]	Udział [%]	L. nieinno.	L. inno.	INV
59	11056.571429 < ACT_MALE_SRPIT36_-POZ109	23,7	10,9	264	82	0,030
76	3352 < ACT_MALE_SRPIT36_-POZ109 ⩽ 11056.571429	16,8	41,5	1 095	221	0,001
85	ACT_MALE_SRPIT36_POZ109 ⩽ 3352	13,6	47,6	1 303	205	0,017
			100,0	2 662	508	0,048

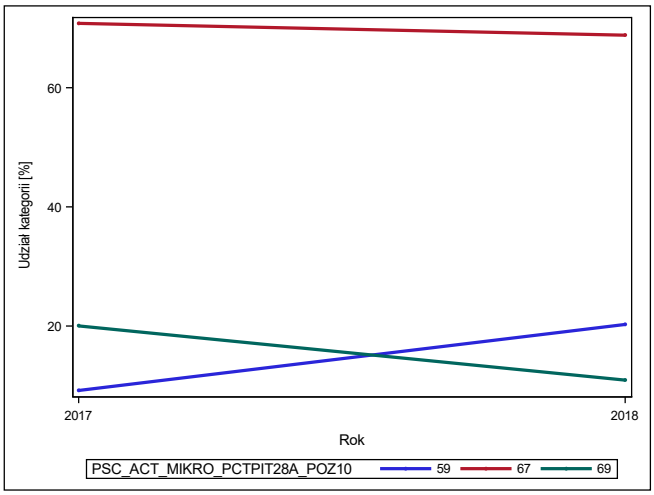


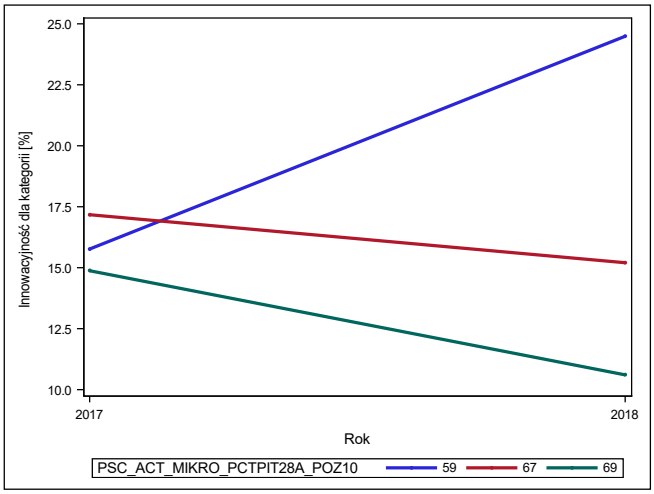


Źródło: opracowanie własne.

Rysunek 2: Kategorie zmiennej ACT_MIKRO_PCTPIT28A_-POZ10

Ocena	Warunek	Inn. [%]	Udział [%]	L. nieinno.	L. inno.	INV
59	ACT_MIKRO_PCTPIT28A_-POZ10 $\leqslant$ 0.1983084509	22,3	14,7	363	104	0,028
67	0.1983084509 $<$ ACT_MIKRO_PCT-PIT28A_POZ10 $\leqslant$ 0.4762347181	15,2	69,9	1 879	338	0,002
69	0.4762347181 $<$ ACT_MIKRO_PCT-PIT28A_POZ10	13,6	15,3	420	66	0,005
			100,0	2 662	508	0,036

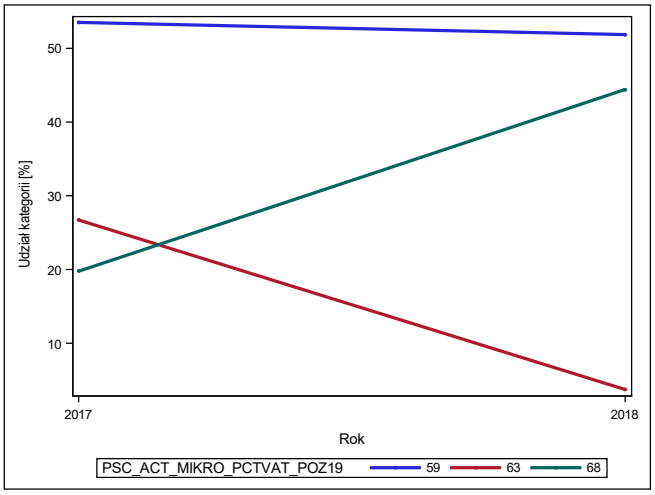


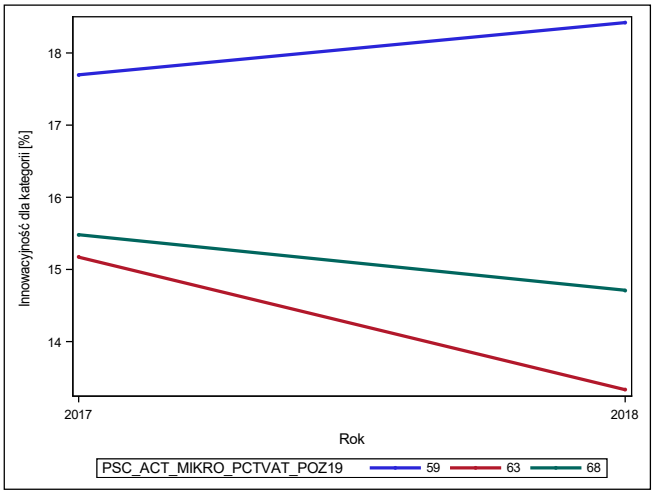


Źródło: opracowanie własne.

Rysunek 3: Kategorie zmiennej ACT_MIKRO_PCTVAT_POZ19

Ocena	Warunek	Inn. [%]	Udział [%]	L. nieinno.	L. inno.	INV
59	ACT_MIKRO_PCTVAT_-POZ19 $\leqslant$ 0.5465141947	17,4	52,9	1 385	291	0,005
63	0.5465141947 $<$ ACT_MIKRO_PCTVAT_-POZ19 $\leqslant$ 0.5572727784	15,6	15,4	412	76	0,000
68	0.5572727784 $<$ ACT_MIKRO_PCTVAT_-POZ19	14,0	31,7	865	141	0,007
			100,0	2 662	508	0,013

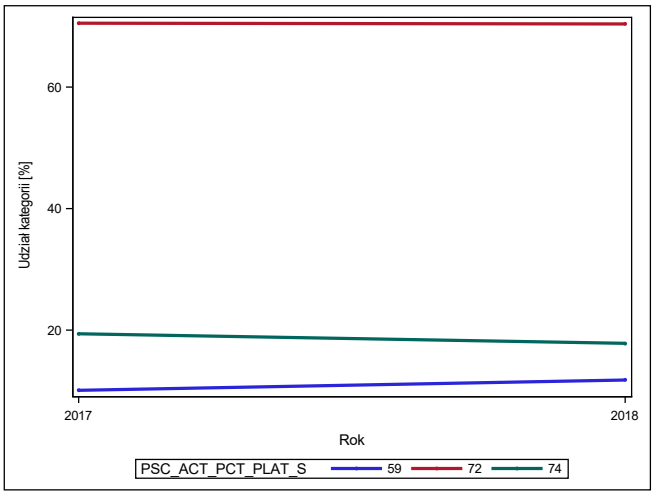


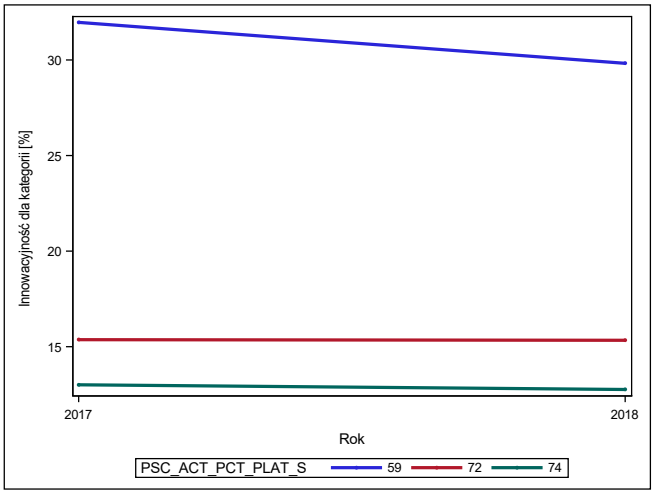


Źródło: opracowanie własne.

Rysunek 4: Kategorie zmiennej ACT_PCT_PLAT_S

Ocena	Warunek	Inn. [%]	Udział [%]	L. nieinno.	L. inno.	INV
59	0.0668769716 $<$ ACT_PCT_PLAT_S	31,3	11,4	248	113	0,113
72	0.0474683544 $<$ ACT_PCT_PLAT_S $\leqslant$ 0.0668769716	14,6	70,2	1 901	324	0,009
74	ACT_PCT_PLAT_S $\leqslant$ 0.0474683544	12,2	18,4	513	71	0,017
			100,0	2 662	508	0,138

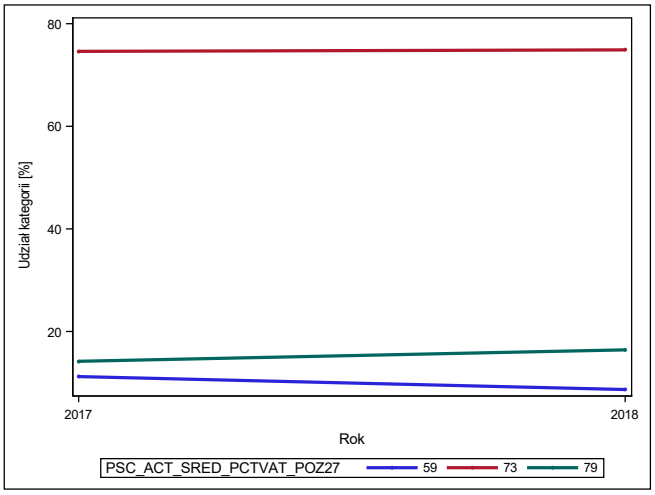


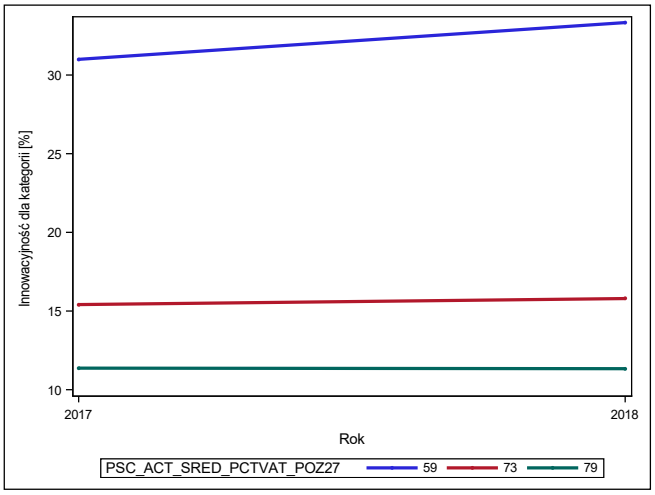


Źródło: opracowanie własne.

Rysunek 5: Kategorie zmiennej ACT_SRED_PCTVAT_POZ27

Ocena	Warunek	Inn. [%]	Udział [%]	L. nieinno.	L. inno.	INV
59	0.0025255184 < ACT_SRED_PCTVAT_-POZ27	31,7	10,1	218	101	0,104
73	0.0000389548 < ACT_SRED_PCTVAT_-POZ27 ⩽ 0.0025255184	15,0	75,3	2 028	359	0,004
79	ACT_SRED_PCTVAT_-POZ27 ⩽ 0.0000389548	10,3	14,6	416	48	0,031
			100,0	2 662	508	0,139





Źródło: opracowanie własne.

2.5. Statystyki zmiennych

Podstawowe statystyki zmiennych wykorzystanych w modelu przedstawione są w tabelach 4-6. Na końcu podrozdziału umieszczony jest słownik oznaczeń statystyk dla zmiennych.

Tabela 4: Statystyki zmiennych, część 1, jakość zmiennych

Zmienna	Typ	P. braków [%]	L. unikalnych	P. mody [%]	Moda
ACT_MALE_SRPIT36_POZ109	INT	0,0	38	33,8	0
ACT_MIKRO_PCTPIT28A_POZ10	INT	0,0	220	13,7	0.315358
ACT_MIKRO_PCTVAT_POZ19	INT	0,0	220	13,7	0.559035
ACT_PCT_PLAT_S	INT	0,0	209	13,7	0.063733
ACT_SRED_PCTVAT_POZ27	INT	0,0	104	13,7	0.001448

Źródło: opracowanie własne.

Tabela 5: Statystyki zmiennych, część 2, predykcyjność i stabilność

Zmienna	Gini [%]	R. Gini [%]	PSI	Intensywność [%]
ACT_MALE_SRPIT36_POZ109	10,9	20,1	0,005707	32,5
ACT_MIKRO_PCTPIT28A_POZ10	8,2	14,1	0,000120	12,5
ACT_MIKRO_PCTVAT_POZ19	5,8	9,8	0,000558	11,3
ACT_PCT_PLAT_S	15,2	29,1	0,001826	18,8
ACT_SRED_PCTVAT_POZ27	15,5	15,4	0,002845	25,0

Źródło: opracowanie własne.

Tabela 6: Statystyki zmiennych, część 3, estymacje modelu regresji logistycznej

Zmienna	Stopnie swobody	Estymacja	Błąd standardowy	Statystyka Walda	P-value
Intercept	1	5.2116	1.0943	22.6818	<.0001
ACT_MALE_SRPIT36_POZ109	1	1.3493	0.2426	30.9443	<.0001
ACT_MIKRO_PCTPIT28A_POZ10	1	0.6165	0.2785	4.9010	0.0268
ACT_MIKRO_PCTVAT_POZ19	1	1.2149	0.5074	5.7334	0.0166
ACT_PCT_PLAT_S	1	0.4577	0.2065	4.9133	0.0267
ACT_SRED_PCTVAT_POZ27	1	0.5069	0.2116	5.7396	0.0166

Źródło: opracowanie własne.

Słownik oznaczeń statystyk

- **Typ** – Skala zmiennej, INT - interwałowa, czyli ciągła (np. wiek), NOM - nominalna, czyli zmienna jakościowa (np. miasto).

- **P. braków** – Udział procentowy braków danych. Brak danych niekoniecznie jest rozumiany jako brak informacji.
- **L. unikalnych** – Liczba unikalnych, różnych, wartości danej zmiennej.
- **P. mody** – Udział procentowy dominanty, mody, czyli wartości występującej najczęściej.
- **Moda** – Wartość mody, dominanty.
- **Gini** – Statystyka Giniego dla zmiennej transformowanej do logit na zbiorze treningowym. Mierzy moc predykcyjną pojedynczej zmiennej, możliwe wartości 0 – 100, patrz (ŘEZÁČ i ŘEZÁČ, 2011).
- **R. Gini** – Różnica względna pomiędzy statystyką Giniego na zbiorze treningowym i walidacyjnym. Mierzy stabilność predykcji, czyli odkrytych reguł decyzyjnych.
- **PSI** – Stabilność zmiennej, odległość pomiędzy rozkładami kategorii zmiennej na zbiorze treningowym i walidacyjnym, liczona statystyką ang. Population Stability Index. Typowe granice $\leqslant 0{,}1$ lub $\leqslant 0{,}5$.
- **Intensywność** – Udział procentowy rozstępu oceny cząstkowej w karcie skoringowej dla zmiennej w całym rozstępnie ocen modelu. Mierzy wpływ zmiennej na finalną ocenę punktową.
- **Stopnie swobody** – Liczba stopni swobody, patrz (Dobson, 2002).
- **Estymacja** – Wartość estymowanego współczynnika regresji w modelu regresji logistycznej, patrz (Dobson, 2002).
- **Błąd standardowy** – Standardowy błąd estymacji, patrz (Dobson, 2002).
- **Statystyka Walda** – Statystyka mierząca istotność zmiennej w modelu regresji logistycznej, patrz (Dobson, 2002).

- **P-value** – Istotność zmiennej liczona przez prawdopodobieństwo krytyczne statystyki Walda, interpretowane jako prawdopodobieństwo hipotezy zerowej, przypuszczenia, że współczynnik regresji może być równy zeru przy zaobserwowanych danych w próbie, czyli nieistotny. Typowa wartość „<.0001” oznacza sytuację bardzo pewną, że hipoteza zerowa może być prawdziwa rzadziej niż raz na 10 tyś. przypadków, patrz (Dobson, 2002).

2.6. Statystyki modelu

Statystyki modelu przedstawione są w tabelach 7 i 8. Ewolucja statystyki Giniego przedstawiona jest w tabeli 9. Na końcu podrozdziału znajduje się słownik oznaczeń statystyk modelu.

Tabela 7: Statystyki modelu - wartości procentowe

Statystyka	Wartość [%]
Gini	24,9
R. Gini	14,2
Gains1	10,2
Gains2	10,2
Gains3	10,2
Gains4	10,2
Gains5	11,4
Gains10	22,0
Gains50	89,0

Źródło: opracowanie własne.

Tabela 8: Statystyki modelu - wartości liczbowe

Statystyka	Wartość
KS Score	0,029697
PSI Score	0,068648
Max VIF	2,285955
Max Pearson	0,733322
Max Con Index	66,341010
Lift1	10,236220
Lift2	5,118110
Lift3	3,412073
Lift4	2,559055
Lift5	2,283465
Lift10	2,204724
Lift50	1,779528

Źródło: opracowanie własne.

Słownik oznaczeń statystyk

- **Gini** – Statystyka Giniego dla modelu na zbiorze treningowym. Mierzy moc predykcyjną pojedynczej całego modelu, możliwe wartości 0 – 100, patrz (ŘEZÁČ i ŘEZÁČ, 2011).
- **R. Gini** – Różnica względna pomiędzy statystyką Giniego na zbiorze treningowym i walidacyjnym. Mierzy stabilność predykcji, czyli odkrytych reguł decyzyjnych w całym modelu.

Tabela 9: Ewolucja mocy predykcyjnej modelu - statystyki Giniego w czasie

Rok	Gini [%]
2017	24,0
2018	23,4

Źródło: opracowanie własne.

- **GainsXX** – Moc identyfikacji złych klientów w pierwszych percentylach próby. Ile procent złych klientów ze wszystkich w próbie (zbiorze treningowym) można zidentyfikować wybierając pierwszy XX-percentyl według ocen punktowych modelu.

- **KS Score** – Statystyka Kołmogorowa–Smirnowa mierząca stabilność rozkładów ocen punktowych pomiędzy zbiorem treningowym i walidacyjnym.

- **PSI Score** – Stabilność ocen punktowych, odległość pomiędzy rozkładami ocen na zbiorze treningowym i walidacyjnym, liczona statystyką ang. Population Stability Index. Typowe granice $\leqslant 0{,}1$ lub $\leqslant 0{,}5$.

- **Max VIF** – Maksymalny współczynnik inflacji wariancji (ang. variance infation factor) dla zmiennych w modelu, mierzący poziom współliniowości zmiennych w modelu regresji logistycznej. Oparty na współczynniku determinacji regresji liniowej liczonym dla modelu wyjaśniającego jedną zmienną względem wszystkich pozostałych. Typowe granice: < 3, < 5 lub < 10, patrz (Ćwik i Koronacki, 2005).

- **Max Pearson** – Maksymalny współczynnik korelacji Pearsona pomiędzy parami zmiennych w modelu, mierzy także poziom współliniowości zmiennych.

- **Max Con Index** – Maksymalny index warunkujący (ang. Condition Index) mierzący poziom wpółliniowości, oparty na wartościach własnych macierzy kowariancji, typowe granice < 50, < 100 lub < 150, patrz (Welfe, 2003).

- **LiftXX** – Podobnie do GainsXX, jest to moc identyfikacji złych klientów w pierwszych percentylach próby. Ile razy wiecej złych klientów w próbie (zbiorze treningowym) można zidentyfikować wybierając pierwszy XX-percentyl według ocen punktowych modelu w stosunku do modelu losowego.

2.7. Kalibracja do wartości prognozy wskaźnika innowacyjności

Kalibracja przekształcająca wartość oceny punktowej w wartość wskaźnika innowacyjności, czyli udziału innowacyjnych firm w gminie (P. inno) jest obliczona na wszystkich dostępnych danych na podstawie funkcji odwrotnego logitu i wyraża się następującym wzorem:

P. inno=1/(1+exp(-(-0.032914522*SCORECARD_POINTS +9.8687751246))).

Wszystkie statystyki historyczne sprawdzające jakość kalibracji są przedstawione w tabeli 10.

Tabela 10: Testowanie kalibracji

Rok	Liczba	Inno [%]	P. inno [%]
Razem	4 836	16,6	16,6
2017	2 418	16,6	16,1
2018	2 418	16,6	17,1

Źródło: opracowanie własne.

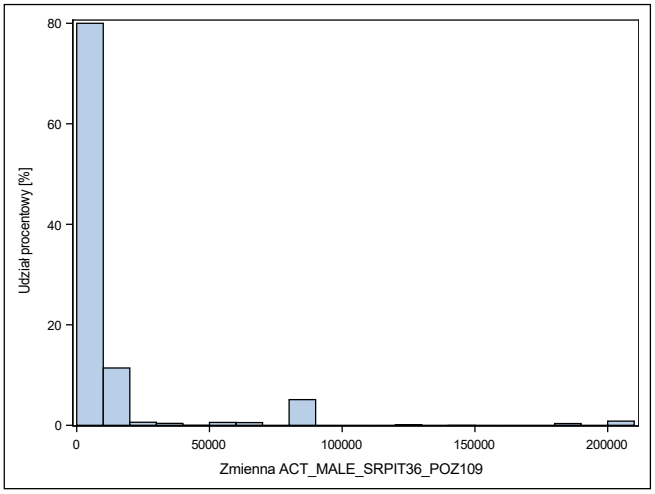
2.8. Analiza jakości zmiennych

Wszelkie histogramy narysowane są tylko dla wartości zakresów zmiennej pomiędzy percentylami P1 i P99, daje to w wielu przypadkach znacząco lepsze możliwości wizualizacji danych. Przedstawione raporty nie wykazują nietypowych wartości, które trzeba byłoby oddzielnie analizować.

Histogramy wraz z percentylami dla każdej ze zmiennych ciągłych w modelu oraz tablicze liczebności z wykresami słupkowymi dla zmiennych nominalnych zostały przedstawione na rysunkach 6-10.

Rysunek 6: Rozkład i statystyki percentyle zmiennej ACT_MALE_-SRPIT36_POZ109

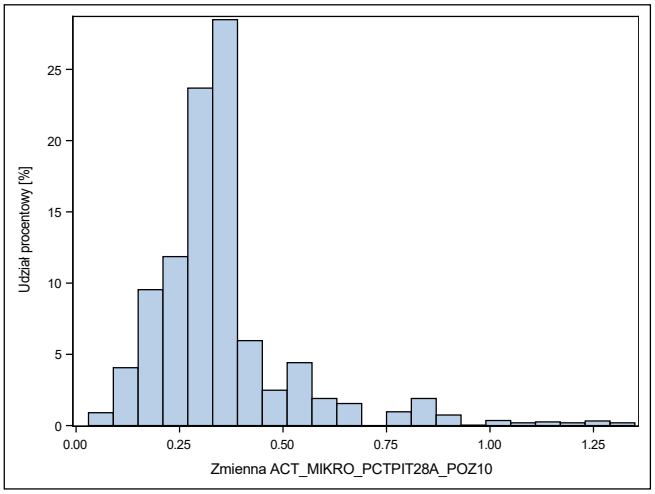
Min	P1	P5	P25	P50	P75	P95	P99	Max
0	0	0	0	3658.75	7182.711	83201.41	223237.1	559380.4



Źródło: opracowanie własne.

Rysunek 7: Rozkład i statystyki percentyle zmiennej ACT_MIKRO_PCTPIT28A_POZ10

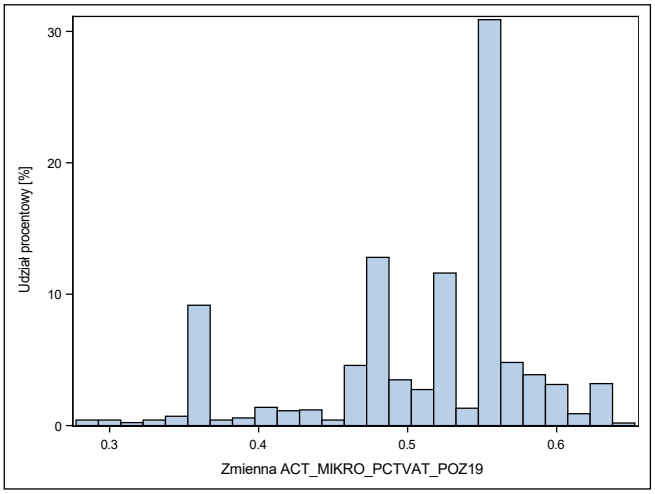
Min	P1	P5	P25	P50	P75	P95	P99	Max
0	0.048641	0.14175	0.262224	0.333028	0.343923	0.818834	1.416493	6.983391



Źródło: opracowanie własne.

Rysunek 8: Rozkład i statystyki percentyle zmiennej ACT_MIKRO_PCTVAT_POZ19

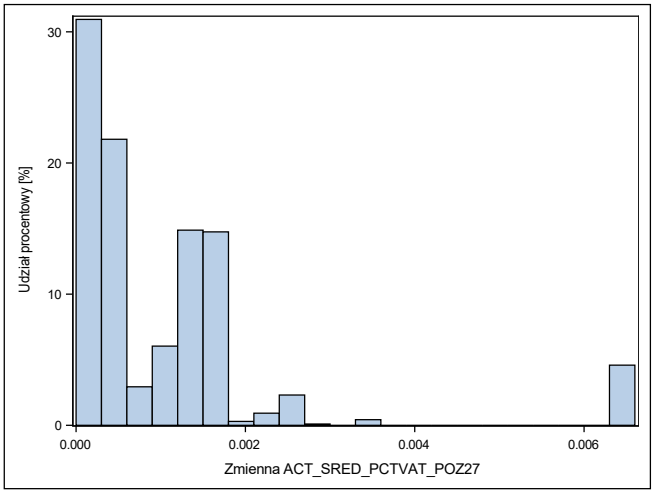
Min	P1	P5	P25	P50	P75	P95	P99	Max
0.162389	0.274519	0.354786	0.480771	0.524126	0.559035	0.613075	0.655948	0.802523



Źródło: opracowanie własne.

Rysunek 9: Rozkład i statystyki percentyle zmiennej ACT_SRED_-PCTVAT_POZ27

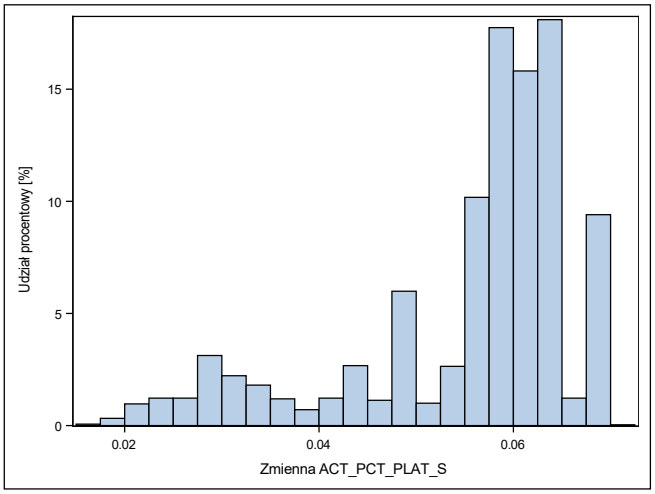
Min	P1	P5	P25	P50	P75	P95	P99	Max
0	0	0	0.000155	0.000495	0.001572	0.006583	0.008002	0.03275



Źródło: opracowanie własne.

Rysunek 10: Rozkład i statystyki percentyle zmiennej ACT_PCT_-PLAT_S

Min	P1	P5	P25	P50	P75	P95	P99	Max
0	0.016393	0.02807	0.051643	0.058438	0.063733	0.06841	0.072543	0.101796



Źródło: opracowanie własne.

Bibliografia

Anderson R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press.

BIS-BASEL (2005). *International convergence of capital measurement and capital standards*. Raport techniczny, Basel Committee on Banking Supervision, Bank For International Settlements. http://www.bis.org. Dostęp: 2012.08.30.

BIS-WP14 (2005). *Studies on validation of internal rating systems, working paper no. 14*. Raport techniczny, Basel Committee on Banking Supervision, Bank For International Settlements. http://www.bis.org. Dostęp: 2012.08.30.

Dobson A.J. (2002). *An Introduction to Generalized Linaer Models*. Chapman & Hall.

ŘEZÁČ M., ŘEZÁČ F. (2011). *How to measure the quality of credit scoring models*. Czech Journal of Economics and Finance, 61 (5), s. 486–507. http://journal.fsv.cuni.cz/storage/1228_rezac.pdf. Dostęp: 2015.02.21.

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

Janc A., Kraska M. (2001). *Credit-scoring: nowoczesna metoda oceny zdolności kredytowej*. Biblioteka Menedżera i Bankowca "Zarządzanie i Finanse".

Koronacki J., Mielniczuk J. (2001). *Statystyka dla studentów kierunków technicznych i przyrodniczych*. WNT, Warszawa.

Lessmanna S., Seowb H.V., Baesenscd B., Thomasd L.C. (2013). *Benchmarking state–of–the–art classification algorithms for credit scoring: A ten–year update*. Credit Scoring Conference CRC, Edinburgh. http://www.business-school.ed.ac.uk/crc/conferences/. Dostęp: 2013.09.23.

Przanowski K. (2014). *Credit Scoring w erze Big-Data*. Oficyna Wydawnicza SGH.

Przanowski K. (2015). *Credit Scoring – studia przypadków procesów biznesowych*. Oficyna Wydawnicza SGH.

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.

Thonabauer G., Nosslinger B. (2004). *Guidelines on Credit Risk Management. Credit Approval Process and Credit Risk Management*. Oesterreichische Nationalbank and Austrian Financial Market Authority.

Verstraeten G., den Poel D.V. (2005). *The impact of sample bias on consumer credit scoring performance and profitability*. Journal of the Operational Research Society, 56, s. 981–992.

Welfe A. (2003). *Ekonometria*. PWE, Warszawa.

Ćwik J., Koronacki J. (2005). *Statystyczne systemy uczące się*. Wydawnictwo Naukowo-Techniczna.

Spis rysunków

1 Kategorie zmiennej ACT_MALE_SRPIT36_POZ109 . 14

2 Kategorie zmiennej ACT_MIKRO_PCTPIT28A_POZ10 15

3 Kategorie zmiennej ACT_MIKRO_PCTVAT_POZ19 . 16

4 Kategorie zmiennej ACT_PCT_PLAT_S 17

5 Kategorie zmiennej ACT_SRED_PCTVAT_POZ27 . . 18

6 Rozkład i statystyki percentyle zmiennej ACT_MALE_SRPIT36_POZ109 25

7 Rozkład i statystyki percentyle zmiennej ACT_MIKRO_PCTPIT28A_POZ10 26

8 Rozkład i statystyki percentyle zmiennej ACT_MIKRO_PCTVAT_POZ19 27

9 Rozkład i statystyki percentyle zmiennej ACT_SRED_PCTVAT_POZ27 28

10 Rozkład i statystyki percentyle zmiennej ACT_PCT_PLAT_S 29

Spis tablic

1 Model prognozy innowacyjności 9

2 Liczebności zbiorów wykorzystanych w estymacji modelu 11

3 Liczebność zbioru wszystkich przedsiębiorstw 11

4 Statystyki zmiennych, część 1, jakość zmiennych . . . 19

5 Statystyki zmiennych, część 2, predykcyjność i stabilność 19

6 Statystyki zmiennych, część 3, estymacje modelu regresji logistycznej 19

7 Statystyki modelu - wartości procentowe 22

8 Statystyki modelu - wartości liczbowe 22

9 Ewolucja mocy predykcyjnej modelu - statystyki Giniego w czasie 23

10 Testowanie kalibracji 24