Wojciech Bijak

Kolegium Analiz Ekonomicznych
Szkoła Główna Handlowa w Warszawie

Ubezpieczeniowy Fundusz Gwarancyjny

Piotr Dziel, Stanisław Garstka, Krzysztof Hrycko

Ubezpieczeniowy Fundusz Gwarancyjny

Metody i modele statystyczne w wykrywaniu nieubezpieczonych posiadaczy pojazdów mechanicznych

Streszczenie

W pracy opisano system stworzony na potrzeby kontroli przeprowadzanej przez Ubezpieczeniowy Fundusz Gwarancyjny w zakresie spełnienia obowiązku posiadania ubezpieczenia odpowiedzialności cywilnej posiadaczy pojazdów mechanicznych. Szczególną uwagę poświęcono zastosowanym algorytmom typowania nieubezpieczonych oraz wykorzystaniu statystycznych systemów uczących się pod nadzorem w procesie wyznaczania najbardziej prawdopodobnych przypadków braku ubezpieczenia. Publikacja zawiera również informacje na temat efektywności rozwiązania oraz jego porównanie z innymi wypracowanymi przez ustawodawcę sposobami kontroli spełnienia opisywanego obowiązku ubezpieczenia.

Słowa kluczowe: kontrola ubezpieczenia OC posiadaczy pojazdów mechanicznych, system automatycznego wykrywania nieubezpieczonych, statystyczne systemy uczące się pod nadzorem

1. Wstęp

Zgodnie z ustawą o ubezpieczeniach obowiązkowych, Ubezpieczeniowym Funduszu Gwarancyjnym i Polskim Biurze Ubezpieczycieli Komunikacyjnych (dalej również: UoUO), spełnienie obowiązku zawarcia umowy ubezpieczenia odpowiedzialności cywilnej posiadaczy pojazdów mechanicznych (dalej również: OC p.p.m.) za szkody powstałe w związku z ruchem tych pojazdów

podlega kontroli wykonywanej przez organy do tego obowiązane lub uprawnione. Wśród podmiotów obowiązanych do kontroli można wymienić m.in. Policję czy organy właściwe w sprawach rejestracji pojazdów. Ubezpieczeniowy Fundusz Gwarancyjny (dalej również: UFG) jest jednym z podmiotów uprawnionych do przeprowadzania wskazanej czynności. Prawne usankcjonowanie takiej możliwości pozwoliło na implementację systemowej kontroli ubezpieczenia OC z wykorzystaniem bazy danych Ośrodka Informacji UFG (dalej również: OI) oraz baz referencyjnych, np. Centralnej Ewidencji Pojazdów Ministerstwa Spraw Wewnętrznych (dalej również: CEP oraz MSW). Kontrole własne UFG, stanowiące wcześniej wyłącznie uzupełnienie działań realizowanych przez podmioty zewnętrzne, stały się aktywnym narzędziem wykrywania nieubezpieczonych. W ten sposób UFG spełnia funkcję kontrolno-represyjną w zakresie weryfikacji spełnienia obowiązku posiadania ubezpieczenia odpowiedzialności cywilnej posiadaczy pojazdów mechanicznych oraz może lepiej realizować funkcję kompensacyjną w odniesieniu do zdarzeń spowodowanych przez nieubezpieczonych posiadaczy pojazdów.

W UFG stworzono automatyczny i działający w cyklach system wykrywający okresy bez ubezpieczenia OC p.p.m. w historii pojazdu. Podstawową rolę w wykrywaniu nieubezpieczonych odgrywają przekazywane z zakładów ubezpieczeń dane, które ze względu na wewnętrzne procesy każdego z ubezpieczycieli charakteryzuje różna jakość i terminowość zasilania OI UFG. Uwzględniając powyższe, należy stwierdzić, że każdego miesiąca system identyfikuje kilkaset tysięcy przypadków braku informacji o ubezpieczeniu OC p.p.m. w bazie OI UFG, z których jedynie mała część dotyczy faktycznie nieubezpieczonych. Reszta zaś wynika z jakości i terminowości danych przekazywanych do OI UFG przez zakłady ubezpieczeń. W tej sytuacji niezbędne jest wykorzystanie modeli statystycznych do odróżniania pojazdów bez ochrony ubezpieczeniowej od tych, które ze względu na jakość danych tylko wydają się jej nie mieć.

W pracy opisano wykorzystywane modele statystyczne, efektywność zastosowanego rozwiązania wyznaczoną na podstawie wstępnych wyników testów modeli oraz wyników działania po wdrożeniu tych modeli. Szczególnie skupiono się na ocenie skuteczności statystycznych systemów uczących się pod nadzorem, w tym modeli regresji logistycznej, drzew decyzyjnych oraz sieci neuronowych. Wskazana skuteczność jest porównywana ze skutecznością działań przeprowadzanych przez podmioty zewnętrzne.

2. Koncepcja systemowego wykrywania nieubezpieczonych posiadaczy pojazdów mechanicznych

2.1. Kontrola ubezpieczenia OC posiadaczy pojazdów mechanicznych

W lutym 2012 r. zaczęła obowiązywać nowelizacja UoUO, która prawnie usankcjonowała kompetencje UFG w zakresie przetwarzania danych zawartych w bazie Ośrodka Informacji w celu kontroli spełnienia obowiązku ubezpieczenia OC p.p.m. (art. 102 ust. 7 UoUO). Kontrole własne prowadzono już wcześniej, a w 2011 r. uruchomiono projekt, którego efektem było stworzenie w pełni automatycznego i działającego w cyklach systemu wykrywania nieubezpieczonych.

Przeprowadzenie systemowej kontroli ubezpieczenia OC p.p.m. przez UFG jest możliwe dzięki wykorzystaniu rejestru umów ubezpieczenia działu II, grupy 10 załącznika do ustawy o działalności ubezpieczeniowej[1]. Informacje dotyczące umów ubezpieczenia są przekazywane przez zakłady ubezpieczeń z zachowaniem terminu 14 dni od momentu zawarcia umowy. W procesach kontroli ubezpieczenia OC p.p.m. są również wykorzystywane dane CEP. Uzupełniają one zasób bazy OI o informacje dotyczące czynności związanych ze zmianą stanu pojazdu, takich jak rejestracja lub wyrejestrowanie, oraz służą uzyskaniu niezbędnych informacji o właścicielach pojazdu.

W UoUO (art. 84 ust. 2) są enumeratywnie wymienione pozostałe organy obowiązane lub uprawnione do kontroli ubezpieczenia OC p.p.m. Wskazane organy po wykonaniu czynności kontrolnych zawiadamiają UFG, w którym prowadzone jest dalsze postępowanie wyjaśniające. W przypadku potwierdzenia braku spełnienia obowiązku ubezpieczenia OC p.p.m. nakładana jest opłata karna, która zależy od następujących czynników:

1) rodzaju pojazdu;
2) okresu pozostawania bez ubezpieczenia w roku kalendarzowym kontroli;
3) minimalnego wynagrodzenia za pracę w roku kontroli.

W przypadku dwóch pierwszych czynników wyróżniono po trzy grupy. Ostatni element zapewnia automatyczną indeksację wysokości opłaty, jako uzależnionej od jednego ze wskaźników ekonomicznych. Tabela 1 zawiera zestawienie

[1] Ośrodek Informacji UFG prowadzi rejestr umów ubezpieczenia działu II, grup 3 i 10 załącznika do ustawy o działalności ubezpieczeniowej z wyłączeniem ubezpieczenia odpowiedzialności cywilnej przewoźnika oraz gromadzi dane dotyczące zdarzeń powodujących odpowiedzialność zakładu ubezpieczeń z tytułu zawartych umów.

wysokości opłaty karnej wyrażonej w jednostkach minimalnego wynagrodzenia za pracę w zależności od pierwszego i drugiego czynnika. Należy zwrócić uwagę na fakt, że znacznie wyższe koszty od wskazanych poniżej może ponieść nieubezpieczony posiadacz pojazdu będący sprawcą wypadku. Jest on bowiem zobowiązany do zwrotu spełnionego przez UFG świadczenia i poniesionych kosztów (art. 110 ust. 1 UoUO).

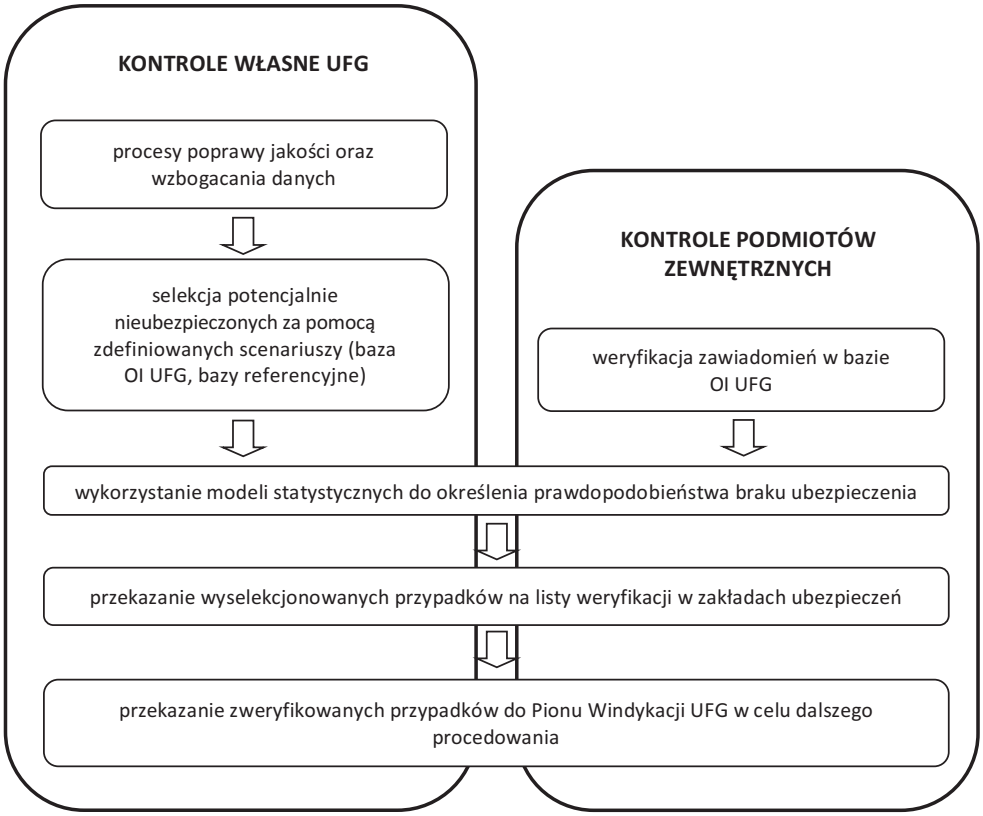
Tabela 1. Opłaty karne za brak ubezpieczenia OC p.p.m. wyrażone w jednostkach minimalnego wynagrodzenia za pracę

Okres pozostawania bez ubezpieczenia	Rodzaj pojazdu		
	samochód osobowy	samochód ciężarowy, ciągnik samochodowy, autobus	pozostałe pojazdy
1–3 dni	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{1}{15}$
4–14 dni	1	$\frac{3}{2}$	$\frac{1}{6}$
Powyżej 14 dni	2	3	$\frac{1}{3}$

Źródło: opracowanie własne na podstawie UoUO.

2.2. Wykrywanie nieubezpieczonych w procesie kontroli własnych UFG

Systemowe wykrywanie nieubezpieczonych jest procesem uruchamianym cyklicznie. Iterację procesu rozpoczyna wyznaczenie potencjalnych przypadków braku ubezpieczenia poddawanych dalszym procedurom weryfikacyjnym, a kończy wysłanie wezwania do wniesienia opłaty karnej. System obejmuje przypadki z kontroli własnych UFG, a docelowo będzie wykorzystywał również zawiadomienia z podmiotów zewnętrznych, obsługiwane obecnie w sposób jednostkowy przez pracowników UFG. W przyszłości każde takie zawiadomienie będzie podlegało procesowi automatycznej weryfikacji pod kątem występowania ubezpieczenia w bazie OI UFG oraz możliwe będzie wykorzystanie modeli statystycznych do początkowej oceny ich zasadności. Rysunek 1 ujmuje zarówno przypadki z kontroli własnych, jak i zawiadomienia z podmiotów zewnętrznych.



Rysunek 1. System wykrywania nieubezpieczonych w procesie kontroli własnych UFG i kontroli podmiotów zewnętrznych

Źródło: opracowanie własne.

Jednym z pierwszych etapów procesu systemowej kontroli jest analiza danych znajdujących się w bazie OI. Do weryfikacji trafiają dane, które wcześniej zostały poddane procesom poprawy jakości oraz są wzbogacone o informacje ze źródeł zewnętrznych. Do źródeł zewnętrznych zalicza się m.in. rejestry państwowe, w tym dane CEP, czy słowniki referencyjne, np. słowniki adresów, marek i modeli samochodowych. Wskazany etap bazuje na scenariuszach, z których każdy odwzorowuje inny sposób powstawania okresu bez ubezpieczenia. Dla przypadków z poszczególnych scenariuszy jest tworzona historia ubezpieczenia, na bazie której są typowane pojazdy oraz ich posiadacze potencjalnie nieposiadający ubezpieczenia OC p.p.m. Równocześnie dla zidentyfikowanych przypadków jest wyznaczane szerokie spektrum zmiennych. Wytypowane przypadki wraz ze zmiennymi opisującymi są umieszczone w tzw. tabeli analitycznej, która jest źródłem danych wejściowych dla stworzonych w procesie uczenia modeli

statystycznych. Na podstawie wartości zmiennych oblicza się prawdopodobieństwo braku ubezpieczenia. Wykorzystanie modeli predykcyjnych pozwala na umieszczenie pojazdów na jednej z list weryfikacji w zakładach ubezpieczeń. Lista weryfikacji obligatoryjnej obejmuje przypadki największego – wyznaczonego na podstawie modelu – prawdopodobieństwa braku ubezpieczenia. Wskazana lista, w odróżnieniu od listy opcjonalnej, w procesie zakładania spraw opłatowych jest traktowana priorytetowo[2].

Weryfikacja przeprowadzana w zakładach ubezpieczeń zapewnia jeszcze wyższą trafność ostatecznego typowania pojazdów oraz ich posiadaczy bez ubezpieczenia OC p.p.m. Ze względu na możliwą aktualizację danych z zakładów ubezpieczeń, w kluczowych momentach procesu odbywa się również ponowna weryfikacja w bazie OI.

3. Modelowanie predykcyjne na potrzeby systemu wykrywania nieubezpieczonych

3.1. Charakterystyka statystycznych systemów uczących się pod nadzorem

Statystyczne systemy uczące się dzieli się najogólniej na systemy uczące się pod nadzorem i systemy uczące się bez nadzoru. Mianem pierwszej grupy określa się modele statystyczne tworzone z wykorzystaniem informacji o przynależności badanych obiektów do określonej klasy. W przypadku takich modeli podstawę budowy funkcji zwanej klasyfikatorem stanowi próba ucząca. Na potrzeby analizy zakładamy, że dysponujemy niezależnymi, prostymi próbami losowymi o liczebnościach n_0 oraz n_1, pobranymi z dwóch różnych populacji. W przedstawionym przypadku mówimy o populacji (zamiennie są stosowane pojęcia grupy lub klasy) ubezpieczonych oraz nieubezpieczonych posiadaczy pojazdów mechanicznych, o których informacja pochodzi z utworzonych wcześniej zawiadomień w systemie opłatowym UFG. Populacje oznaczamy odpowiednio przez $Y = 0$, $Y = 1$.

[2] Sprawa opłatowa odnosi się do utworzonego w systemie informatycznym Pionu Windykacji UFG wezwania do uiszczenia opłaty karnej w związku z brakiem posiadania ubezpieczenia OC p.p.m. W sprawie opłatowej są zawarte szczegółowe informacje dotyczące podmiotu zobowiązanego oraz pojazdu, w którego przypadku powstała nieciągłość w ochronie ubezpieczeniowej.

3.1.1. Uogólnione modele liniowe

W przypadku uogólnionych modeli liniowych (ang. *Generalized Linear Model* – GLM) przyjmuje się, że rozkład zmiennej zależnej należy do tzw. wykładniczej rodziny rozkładów. W GLM wartość oczekiwaną zmiennej objaśnianej prognozuje się za pomocą odwrotności tzw. funkcji łączącej (ang. *link function*).

Uogólniony model liniowy można zapisać w postaci:

$$Y \mid \boldsymbol{X} \sim F\left(\theta\right),$$

$$g\left(E\left(Y|\boldsymbol{X}\right)\right) = g\left(\mu\right) = \beta_0 x_0 + \beta_1 x_1 + \ldots + \beta_q x_q = \boldsymbol{X}^T \boldsymbol{\beta},$$

gdzie Y to zmienna objaśniana, $\boldsymbol{X}$ to wektor zmiennych objaśniających, $g\left(\cdot\right)$ jest funkcją łączącą, $\boldsymbol{\beta}$ stanowi wektor poszukiwanych parametrów, a obserwacji podlegają wartości zmiennej losowej o rozkładzie F indeksowanej parametrem θ. W omawianym przypadku zmienna objaśniana przyjmuje wyłącznie wartości ze zbioru $\left\{0,1\right\}$, a zatem warunkowy rozkład ($Y \mid \boldsymbol{X}$) to rozkład dwumianowy z prawdopodobieństwem sukcesu p (oznaczenie $B\left(1,p\right)$). W przypadku regresji logistycznej jest wykorzystywana funkcja łącząca postaci:

$$g\left(p\right) = \ln\left(\frac{p}{1-p}\right).$$

Modelowane jest prawdopodobieństwo wystąpienia sukcesu p (prawdopodobieństwo realizacji zdarzenia oznaczanego jako 1 dla obserwowanych wartości wektora $\boldsymbol{X}$ – prawdopodobieństwo braku ubezpieczenia):

$$p = P\left(Y = 1|\boldsymbol{X} = \boldsymbol{x}\right) = E\left(Y \mid \boldsymbol{X} = \boldsymbol{x}\right) = \frac{exp\left(\boldsymbol{x}^T\boldsymbol{\beta}\right)}{1+exp\left(\boldsymbol{x}^T\boldsymbol{\beta}\right)}.$$

Wektor parametrów $\boldsymbol{\beta}$ modelu regresji logistycznej estymuje się metodą największej wiarogodności, wykorzystując algorytm iteracyjnie ważonych najmniejszych kwadratów. Z kolei dobór zmiennych objaśniających do modelu może odbyć się za pomocą metod selekcji: krokowej postępującej (ang. *forward selection procedure*), krokowej wstecznej (ang. *backward elimination procedure*), modyfikacji metody selekcji postępującej (ang. *stepwise procedure*) oraz innych[3].

[3] D.T. Larose, *Data Mining Methods and Models*, John Wiley & Sons, Hoboken 2006, s. 123–127.

3.1.2. Drzewa decyzyjne

Drzewa decyzyjne, inaczej klasyfikacyjne, można przedstawić graficznie za pomocą grafu skierowanego, acyklicznego i spójnego. Drzewa zatem mają postać diagramu z wierzchołkami i krawędziami, nazywanymi odpowiednio węzłami oraz gałęziami. Każdy wierzchołek drzewa stanowi podział elementów próby[4]. Ścieżka składająca się z węzłów i gałęzi wskazuje, do której grupy jest kwalifikowany przypadek, a przynależność jest określana na podstawie liczności grup w węźle końcowym, nazywanym również liściem. Przypadek jest klasyfikowany do grupy najliczniej reprezentowanej przez próbę uczącą w ostatnim wierzchołku. W sytuacji, gdy zagadnienie klasyfikacji dotyczy wyłącznie dwóch grup, iloraz liczby przypadków, dla których zdarzenie się zrealizowało, do ogólnej liczby elementów węzła końcowego może być traktowany jako oszacowanie prawdopodobieństwa wystąpienia badanego zdarzenia dla wskazanej ścieżki.

Budowa drzewa klasyfikacyjnego polega na wyborze zasady podziału elementów próby uczącej w poszczególnych węzłach. Najbardziej pożądany podział to taki, w którego przypadku obserwacje w węźle są w miarę jednorodne (większość obserwacji należy do jednej grupy). Przeprowadzenie podziału wymaga zdefiniowania miary niejednorodności klas w węźle oraz miary różnicy między niejednorodnością klas w danym węźle i niejednorodnością klas w węzłach dzieciach. Wśród wskazywanych w literaturze miar różnorodności klas w węźle można wyróżnić miary oparte na: proporcji błędnych klasyfikacji, indeksie Giniego oraz funkcji entropii[5]. Z kolei najpopularniejsze metody budowy drzew klasyfikacyjnych to algorytmy: CART, QUEST, C4.5, C5.0, CHAID.

W przypadku drzew decyzyjnych należy pamiętać o tzw. efekcie przetrenowania (ang. *overfitting*), czyli sytuacji, gdy drzewo ze względu na zwiększanie liczby liści doskonale klasyfikuje obiekty z próby uczącej, a coraz słabiej klasyfikuje obserwacje z próby testowej. W celu uniknięcia takiego zjawiska stosuje się tzw. reguły stopu lub przycinania drzewa (ang. *prunning*)[6].

---

[4] J. Koronacki, J. Ćwik, *Statystyczne systemy uczące się*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008, s. 129–131.

[5] M. Skorzybut, M. Krzyśko, T. Górecki, W. Wołyński, *Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*, Wydawnictwa Naukowo-Techniczne, Warszawa 2008, s. 160–161.

[6] Ibidem, s. 169–171.

3.1.3. Sieci neuronowe

Sieć neuronowa składa się z neuronów powiązanych ze sobą za pomocą siatki połączeń. W każdym z neuronów zostaje wyznaczona wartość, która następnie jest modyfikowana wyznaczoną w algorytmie uczenia wagą. Wagi są modyfikowane w trakcie procesu uczenia, tak by minimalizować różnicę między wzorcem (rzeczywistą wartością zmiennej objaśnianej) a sygnałem wyjściowym, który jest uzyskiwany po przekształceniu za pomocą tzw. funkcji aktywacji (ang. *activation function*).

Wykorzystywane sieci neuronowe zazwyczaj mają budowę wielowarstwową. Wśród warstw wyróżnia się warstwy wejściową i wyjściową oraz tzw. warstwy ukryte. Warstwa wejściowa służy wprowadzaniu wartości zmiennych obserwowanych do sieci, z kolei warstwa wyjściowa wskazuje na wynik obliczeń. Warstwa ukryta stanowi element sieci, którego nie można bezpośrednio obserwować od strony wejścia ani od strony wyjścia, a składa się ona z neuronów, które przetwarzają informacje z warstwy wejściowej lub poprzedniej warstwy ukrytej i przekazują do następnej warstwy ukrytej lub do warstwy wyjściowej[7].

3.2. Kryteria oceny efektywności modeli statystycznych

Modelowanie zjawiska braku ubezpieczenia dla przypadków wytypowanych z wykorzystaniem ustalonych scenariuszy odbyło się za pomocą wskazanych w podpunkcie 3.1 modeli predykcyjnych. Wyboru najlepszego modelu dokonano na podstawie zdefiniowanych kryteriów. W zależności od tego, czy wykorzystywane były modele parametryczne, zastosowanie miały również standardowe procedury oceny, uwzględniające istotność oszacowania parametrów, analizę reszt, miary dopasowania czy wartości kryteriów informacyjnych. Ważną rolę odgrywała również interpretacja merytoryczna parametrów modelu.

Oceny modelu statystycznego można dokonać za pomocą wskazanych poniżej kryteriów[8]. Ze względu na specyfikę procesu wykrywania nieubezpieczonych do oceny jakości modelu są wykorzystywane wyłącznie niektóre ze wskazanych poniżej miar:

---

[7] R. Tadeusiewicz, *Sieci neuronowe*, Akademicka Oficyna Wydawnicza, Warszawa 1993, s. 12–13.

[8] Por. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statisical Learning: Data Mining, Inference, and Prediction*, Springer, New York 2009, s. 233, 241–245, 317; *Zaawansowane metody analiz statystycznych*, red. E. Frątczak, Oficyna Wydawnicza SGH, Warszawa 2012, s. 562–566.

1. Macierz błędów

Przedstawiona w tabeli 2 macierz błędów w wierszach ma rzeczywiste wskazania grupy, natomiast w kolumnach wskazania grupy z modelu. Łączny błąd skonstruowanego klasyfikatora nazywa się błędem klasyfikacji i wyraża się wzorem:

$$e\left(\hat{d}\right) = P\left(\hat{d}\left(\boldsymbol{X}\right) \neq Y | L_n\right),$$

gdzie $\hat{d}\left(\boldsymbol{X}\right)$ oznacza regułę klasyfikującą obserwację $\boldsymbol{X}$, a para losowa $\left(\boldsymbol{X}, Y\right)$ jest niezależna od próby uczącej L_n. Oceny tego prawdopodobieństwa można dokonać również za pomocą próby uczącej. Ze względu na proces uczenia przeprowadzany na wskazanej próbie, do oceny zaleca się wykorzystanie próby testowej. Kwalifikacja wytypowanych przypadków do określonej grupy odbywa się na podstawie otrzymanej wartości prawdopodobieństwa z modelu. Jednym ze sposobów określenia wartości progowej, która rozdziela obserwacje należące do grupy przypadków zasadnych i niezasadnych, jest kierowanie się zasadą minimalizacji prawdopodobieństwa błędnej kwalifikacji. Według wskazanego podejścia, najlepszy model to taki, który kwalifikuje największą liczbę przypadków do prawdziwej grupy.

Tabela 2. Macierz błędów

Prawdziwa grupa	Wskazania z modelu		Razem
	$\hat{d}\left(\boldsymbol{X}\right)=0$	$\hat{d}\left(\boldsymbol{X}\right)=1$	
$Y = 0$	true negative (TN)	false positive (FP)	TN+FP
$Y = 1$	false negative (FN)	true positive (TP)	FN+TP

Źródło: opracowanie własne.

2. Krzywa ROC (ang. *Receiver Operating Characteristics Curve*)

Wartości krzywej wyznacza się dla $t \in \left[0,1\right]$. Obserwacja jest klasyfikowana do grupy 1, gdy prawdopodobieństwo *a posteriori* tej przynależności spełnia warunek $P\left(\hat{d}\left(\boldsymbol{X}\right)=1\right) > t$. W ten sposób dla poszczególnych wartości $t \in \left[0,1\right]$ jest wyznaczana macierz błędów. Estymator krzywej jest postaci:

$$ROC\left(\cdot\right) = \left\{\left(\alpha\left(t\right), \beta\left(t\right)\right) : t \in \left[0,1\right]\right\},$$

gdzie

$$\alpha\left(t\right) = P\left(\hat{d}\left(\boldsymbol{X}\right)=1 | Y=0\right) = \frac{FP}{TN+FP},$$

$$\beta(t) = P\left(\hat{d}(\boldsymbol{X}) = 1 | Y = 1\right) = \frac{TP}{FN + TP}.$$

Modele silnie dyskryminujące grupy mają krzywe ROC leżące blisko wierzchołka (0,1) kwadratu jednostkowego. Dodatkowo, prawdopodobieństwo $P(\hat{d}(\boldsymbol{X}) = 0 | Y = 0)$ nazywamy specyficznością, z kolei prawdopodobieństwo $P(\hat{d}(\boldsymbol{X}) = 1 | Y = 1)$ to czułość testu opartego na hipotezie H_0 (obserwacja pochodzi z populacji 0) względem hipotezy H_1 (obserwacja pochodzi z populacji 1). Na krzywą ROC można zatem spojrzeć jak na wykres mocy (czułości) względem błędu pierwszego rodzaju (1 – specyficzność) wskazanego wyżej testu[9]. Wykorzystywanym wskaźnikiem jest również pole pod krzywą (ang. *Area Under Curve* – AUC). Im większa jest wartość tego pola, tym lepiej dany klasyfikator odróżnia przypadki zasadne od niezasadnych.

3. Krzywa skumulowanego „podbicia" (ang. *Cumulative Lift Curve*)

Wartości krzywej informują o tym, ile razy więcej przypadków zasadnych znajduje się w kolejnych skumulowanych percentylach próby z najwyższym oszacowaniem prawdopodobieństwa przynależności do grupy 1 w porównaniu z typowaniem losowym. Wskaźnik ten jest wyznaczany najczęściej dla kilku pierwszych percentyli badanej próby, a wynika to z kierowania zawiadomień wyłącznie do potencjalnie zobowiązanych z najwyższym oszacowaniem prawdopodobieństwa braku ubezpieczenia. Im większa jest wartość wskaźnika skumulowanego „podbicia" dla określonego percentyla, tym reguła klasyfikacyjna lepiej dyskryminuje przypadki z dwóch różnych populacji. Pod uwagę należy również brać fakt, że maksymalna wartość wskaźnika zależy od proporcji tzw. 1 w zbiorze. Maksymalna wartość wskaźnika, która może być uzyskana, jest równa odwrotności tej proporcji.

4. v-krokowa metoda sprawdzenia krzyżowego (zamiennie używana jest nazwa – v-krotna kroswalidacja)

Metoda polega na podziale zbioru uczącego na v podzbiorów, przy czym $(v-1)$ z nich tworzy próbę uczącą, natomiast pozostały zbiór stanowi próbę testową. Procedura jest powtarzana v razy dla każdego z podzbiorów testowych. Dla tak skonstruowanych klasyfikatorów jest wyznaczany błąd klasyfikacji. Należy pamiętać o tym, że ze względu na wykorzystanie różnych prób uczących mogą być w rzeczywistości konstruowane modele, których parametry będą się

[9] J. Ćwik, J. Mielniczuk, *Statystyczne systemy uczące się. Ćwiczenia w oparciu o pakiet R*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2009, s. 63.

różnić, i dlatego uważa się, że testowaniu w tym przypadku w większym stopniu podlega metoda statystyczna niż sam model.

5. Kryteria informacyjne

Wśród nich najpopularniejsze to bayesowskie kryterium informacyjne (ang. *Bayesian Information Criterion* – BIC), kryterium informacyjne Akaikego (ang. *Akaike Information Criterion* – AIC), kryterium Hannana–Quinna czy Shibaty. Kryteria te uwzględniają miarę dopasowania modelu oraz liczbę wykorzystanych parametrów. Model najlepszy to taki, dla którego wartość kryterium jest najmniejsza.

6. Test Kołmogorowa–Smirnova

Wartości prawdopodobieństwa zajścia badanego zdarzenia wyznaczone na podstawie utworzonego modelu statystycznego są traktowane jako zmienna, która następnie jest analizowana w poszczególnych, rzeczywistych grupach przynależności. Hipoteza zerowa, mówiąca o braku istotności różnic w rozkładach, jest odrzucana na poziomie istotności α, gdy

$$D_{n_0,n_1} > c(\alpha),$$

gdzie

$$D_{n_0,n_1} = \sqrt{\frac{n_0 n_1}{n_0 + n_1}} \sup_x \left| F_{n_0}(x) - F_{n_1}(x) \right|.$$

$F_{n_0}(x)$, $F_{n_1}(x)$ oznacza dystrybuantę empiryczną opartą na próbach o licznościach odpowiednio n_0 oraz n_1, z kolei $c(\alpha)$ można wyznaczyć ze wzoru $\alpha = 1 - H(c)$, gdzie H oznacza dystrybuantę rozkładu Kołmogorowa. Ten model, dla którego p-wartość testu jest najmniejsza, jest uznawany za najlepiej dyskryminujący przypadki z dwóch grup. Przy założeniu, że próba ucząca i testowa dla poszczególnych modeli jest jednakowa, alternatywnie można posługiwać się poniższą statystyką i wybrać model, dla którego jej wartość jest największa:

$$\sup_x |F_{n_0}(x) - F_{n_1}(x)|.$$

3.3. Algorytmy typowania nieubezpieczonych

Systemowe wykrywanie nieubezpieczonych pojazdów oraz ich posiadaczy jest wykonywane przy zastosowaniu określonych algorytmów wyznaczania

przypadków podejrzanych. Obecnie są stosowane cztery algorytmy, odwzorowujące scenariusze powstawania okresu bez ubezpieczenia w historii pojazdu. Dwa pierwsze dotyczą wskazania okresu bez ubezpieczenia *post factum*, dwa kolejne to algorytmy prewencyjne, mające na celu typowanie braku ubezpieczenia również w momencie kontroli. Poszczególne algorytmy różnicuje występowanie okresu bez ubezpieczenia, który może być:

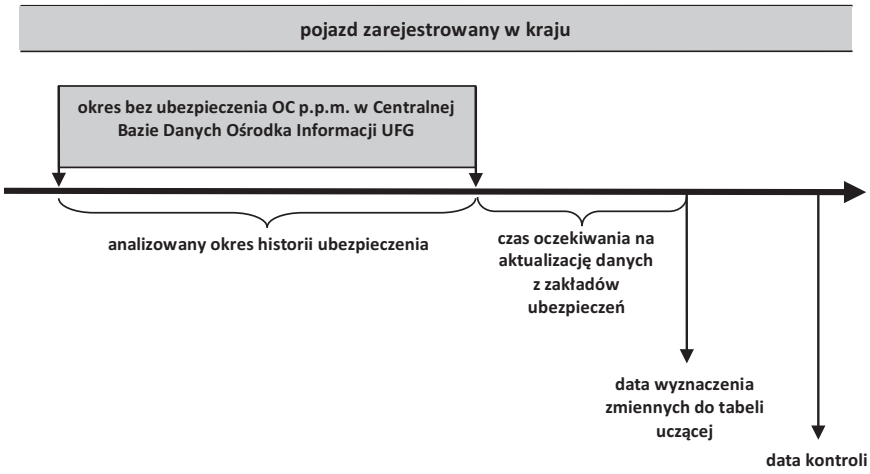
1) ograniczony, a zatem występuje ochrona ubezpieczeniowa zarówno przed, jak i po wskazanym okresie;
2) ograniczony i istnieje od momentu pierwszej rejestracji w kraju do ostatniego dnia przed pierwszym dniem ochrony ubezpieczeniowej;
3) nieograniczony (trwa również w momencie przeprowadzania kontroli przez UFG) oraz istnieje historia ubezpieczeniowa pojazdu w bazie OI;
4) nieograniczony (trwa również w momencie przeprowadzania kontroli przez UFG) oraz nie istnieje historia ubezpieczeniowa pojazdu w bazie OI.

Ze względu na wolumen danych uzyskanych dla procesów uczenia, ale również na poziom wypełnienia zmiennych dla przypadków, gdy nie istnieje historia ubezpieczeniowa pojazdu lub jest ona bardzo krótka, utworzenie modeli statystycznych było możliwe wyłącznie dla algorytmów 1 i 3. W pozostałych przypadkach jest wykorzystywany system reguł eksperckich, który wybranym czynnikom, zidentyfikowanym jako mające wpływ na ryzyko braku ubezpieczenia, nadaje określoną wagę. Należy podkreślić fakt, że ze względu na dostęp UFG do baz referencyjnych MSW możliwe jest wykorzystanie algorytmów 2–4, bez których precyzyjne określenie okresu obowiązku ubezpieczenia byłoby utrudnione.

Na rysunku 2 zostały przedstawione przykładowe zależności czasowe dla algorytmu 4. Parametr „czas oczekiwania" oznacza minimalny czas, który musi upłynąć od pierwszego dnia okresu bez ubezpieczenia do momentu kwalifikacji przypadku jako potencjalny brak ubezpieczenia. Jest on wymagany ze względu na obowiązujące zakłady ubezpieczeń terminy dotyczące zasilenia bazy OI. Parametr „analizowany okres" oznacza czas, dla którego jest przeprowadzana analiza historii ubezpieczeniowej w bazie OI. Moment wyznaczenia zmiennych do tabeli uczącej poprzedza okres wystąpienia sprawy opłatowej w systemie merytorycznym, co ma na celu uniknięcie wpływu działań posiadacza, które mogłyby wystąpić po dacie kontroli.

Algorytmy wykrywania nieubezpieczonych bazują w dużej mierze na wytworzonych algorytmach łączących pojazdy w klastry. Dzięki nim możliwe jest wyznaczenie okresów bez ubezpieczenia również w przypadku pomyłek edycyjnych w identyfikatorach czy zmiany numerów rejestracyjnych. Jeden obiekt fizyczny

w bazie danych może być zapisany w różnych rekordach, a celem łączenia jest uzyskanie kompletnej informacji na jego temat[10].



Rysunek 2. Zależności czasowe dla algorytmu typowania nieubezpieczonych, gdy okres bez ubezpieczenia trwa również w momencie przeprowadzania kontroli przez UFG oraz nie istnieje historia ubezpieczeniowa pojazdu w bazie OI

Źródło: opracowanie własne.

3.4. Opis i przygotowanie danych wykorzystanych w modelowaniu

Budowa modeli statystycznych w procesach uczenia z nadzorem wymaga przygotowania zbiorów uczących oraz testowych. Sposób wyznaczania tych zbiorów powinien w jak największym stopniu odzwierciedlać rzeczywiste procesy typowania nieubezpieczonych, funkcjonujące już po wdrożeniu systemu. W szczególności dotyczy to parametryzacji procesu. Spełnienie ustalonych założeń, ale również korzystanie z oddzielnych systemów w procesach wyznaczania zasadności oraz okresów bez ubezpieczenia powoduje, że liczba przypadków uczących oraz testowych jest ograniczona. W odniesieniu do modelowanego zjawiska informacja o zasadności przypadków została pozyskana z systemu merytorycznego obsługującego zawiadomienia z podmiotów zewnętrznych oraz sprawy opłatowe. Jako sprawy opłatowe zasadne zostały oznaczone przypadki, wobec których w toku kontroli nie przedstawiono dokumentów potwierdzających

[10] W. Bijak, K. Hrycko, Ł. Pietrowski, *Sposób na nieubezpieczonych*, „Miesięcznik Ubezpieczeniowy" 2012, kwiecień, s. 16–17.

istnienie ochrony ubezpieczeniowej oraz nie stwierdzono braku takiego obowiązku (zmienna celu przyjmuje wartość 1). Jako sprawy bezzasadne (niezasadne) oznaczono takie, w odniesieniu do których skierowane roszczenie okazało się nieuzasadnione (zmienna celu przyjmuje wartość 0). Docelowo, wraz ze zbiorem wskazującym zasadność przypadków wyznaczono zmienne opisujące badane obiekty. Wartości zmiennych dla określonych przypadków zostały wyznaczone na miesiąc kontroli, odpowiadający pierwszemu miesiącowi wystąpienia sprawy opłatowej w systemie merytorycznym.

Do celów modelowania wyznaczono zmienne pojazdocentryczne oraz podmiotocentryczne, a zatem takie, w których obiektem opisywanym jest pojazd lub podmiot. Wyznaczone zmienne dotyczyły takich obszarów, jak:

1) właściwości techniczne oraz stan techniczny pojazdu;
2) cechy posiadacza pojazdu;
3) historia ubezpieczenia pojazdu oraz posiadacza, a także przebieg ubezpieczenia;
4) wskaźniki jakości danych przesyłanych w odniesieniu do wskazanych obiektów przez zakłady ubezpieczeń;
5) informacje z systemów merytorycznych oraz źródeł zewnętrznych.

Łącznie zdefiniowano ok. 700 zmiennych, które wraz ze zmienną celu zostały umieszczone w zbiorach uczących wyznaczanych dla parametryzowanych okresów. Przygotowano również zbiory testowe, które sprawdzały efektywność stworzonych modeli statystycznych. W jednym i drugim przypadku proporcje spraw zasadnych i niezasadnych zostały zachowane na takim samym poziomie, a ewentualna modyfikacja tych proporcji następowała dla poszczególnych ścieżek modelowania. Liczność przypadków wykorzystanych w procesach uczenia oraz testów przedstawia tabela 3[11].

Wykorzystany poziom nasycenia zjawiska w procesie uczenia modelu dla algorytmu 1 i 3 wynosi odpowiednio ok. 30% i 40% badanej próby. W cyklicznych pętlach procesu dla algorytmów 1–4 jest wyznaczanych maksymalnie do 800 tys. pojazdów oraz ich nieubezpieczonych posiadaczy. Z kolei szacunki UFG wskazują, że nieubezpieczonych posiadaczy jest od 100 tys. do 250 tys. Szacowany zatem poziom nasycenia zjawiska w skrajnie maksymalnym przypadku wynosi niewiele powyżej 30%, co jest zbieżne z wykorzystywanymi poziomami nasycenia w zbiorach uczących oraz testowych.

[11] Liczności wskazane w tabeli dotyczą modeli historycznych wykorzystywanych w procesach wykrywania nieubezpieczonych. Informacja o zmiennych wykluczonych, przykładowych modelach czy wskaźnikach efektywności odnosi się do tych samych modeli.

Tabela 3. Liczność zbiorów wykorzystanych na potrzeby modelowania zjawiska braku ubezpieczenia

Wykorzystany algorytm/zbiór	Liczba spraw opłatowych	Liczba spraw opłatowych zasadnych	Nasycenie zjawiska
1/zbiór uczący	9797	2798	28,6%
1/zbiór testowy	1319	369	28,0%
3/zbiór uczący	5421	2093	38,6%
3/zbiór testowy	1105	418	37,8%

Źródło: opracowanie własne.

Kolejnym etapem przygotowania danych jest wykluczenie zmiennych charakteryzujących się niestabilnością w czasie oraz niskim poziomem wypełnialności. W celu zbadania zjawiska niestabilności zmiennych ciągłych pomiędzy kolejnymi cyklami wykorzystano test Kołmogorowa–Smirnova, sprawdzający, czy rozkłady prawdopodobieństwa zmiennych różnią się istotnie. Liczba przeprowadzonych testów jest uzależniona od cykli, dla których wyznaczono zbiory uczące. Z kolei uwzględnienie kryterium wypełnialności ma na celu ograniczenie liczby zmiennych, dla których poziom wypełnienia podaje w wątpliwość ich możliwości predykcyjne. Analiza jest przeprowadzana dla poziomu wypełnienia i częstotliwości występowania poszczególnych wartości. W tabeli 4 przedstawiono liczbę zmiennych wykluczonych z dalszych analiz ze względu na wskazany warunek.

Przygotowanie danych obejmuje również wiele innych procesów, a w przypadku wykrywania nieubezpieczonych wszystkie z wymienionych poniżej są realizowane na poziomie tworzenia poszczególnych modeli statystycznych:

1) uzupełnienie braków danych;
2) wykrycie obserwacji odstających oraz sposób postępowania z nimi;
3) działanie w przypadku występowania zależności między zmiennymi objaśniającymi.

Metoda uzupełnienia (imputacji) brakujących danych może rzutować na efektywność modelu, a wykorzystanie odpowiedniej metody zależy w dużym stopniu od rodzaju posiadanych danych. Obserwacje odstające mają mniejszy wpływ na modele drzew decyzyjnych niż na modele regresji. Wskazane powyżej przykłady pokazują, że niezbędne jest dalsze przygotowanie danych do modelu. Najczęściej następuje to na poziomie określonej ścieżki modelowania i jest zależne od doświadczeń analityka w tym zakresie.

Tabela 4. Wykluczenie zmiennych ze względu na kryteria wypełnialności

Kryterium	Algorytm 1	Algorytm 3
Liczba wykluczonych zmiennych numerycznych		
95% wypełnienia jedną wartością	1	1
90% wypełnienia przez braki danych	44	59
98% wypełnienia przez 0 lub braki danych	227	49
Liczba wykluczonych zmiennych tekstowych		
95% wypełnienia jedną wartością	0	33
90% wypełnienia przez braki danych	0	31

Źródło: opracowanie własne.

4. Metody i modele statystyczne w wykrywaniu nieubezpieczonych posiadaczy pojazdów mechanicznych

4.1. Wybrane modele regresji logistycznej, drzew decyzyjnych oraz sieci neuronowych

Dla algorytmów 1 i 3 typowania nieubezpieczonych posiadaczy pojazdów oraz dla każdego z opisanych w punkcie 3 modeli predykcyjnych wyznaczono po kilkadziesiąt ścieżek modelowania, zakończonych wyznaczeniem postaci modelu. Weryfikacja modelu i dalsze działania mające na celu poprawę efektywności odbywały się iteracyjnie.

Ostatecznie wdrażane modele predykcyjne są budowane na maksymalnie kilkunastu zmiennych. Na podstawie tworzonych modeli można zaobserwować, że wykorzystywane zmienne dotyczą:

1) zależności czasowych na umowie ubezpieczenia, np. średniej długości okresu bez ubezpieczenia dla analizowanego przedziału czasowego, rzeczywistej długości okresu świadczonej ochrony dla ostatniej umowy ubezpieczenia, zależności między początkiem okresu ochrony a datą wystawienia polisy;
2) wskaźników jakości danych, np. liczby rekordów błędnych do liczby rekordów poprawnych przesłanych przez zakład ubezpieczeń w odniesieniu do pojazdu;
3) informacji o braku ubezpieczenia w okresach wcześniejszych.

W tabeli 5 przedstawiono model regresji logistycznej wykorzystywany do typowania nieubezpieczonych w jednym z wcześniejszych okresów. Model dotyczył przypadków wyznaczonych na podstawie algorytmu 1.

Tabela 5. Oszacowanie modelu regresji logistycznej – algorytm 1

Zmienna	Poziom zmiennej	Ocena parametru	Błąd standardowy	Chi-kwadrat Walda	Pr > Chi-kwadrat
CAR_POL_INS_ALL_FLG_V2	0	0,2467	0,0457	29,12	<,0001
CAR_OWNER_CNT_V12	01: low-2,5	–0,1908	0,0588	10,54	0,0012
CAR_POL_SEND_ALL_DY_V	01: low-30,5, MISSING	–0,15	0,0683	4,83	0,028
CAR_POL_SEND_ALL_DY_V	02:30,5–170,5	0,2368	0,0726	10,62	0,0011
CAR_POL_SEND_ALL_DY_V	03:170,5–780,5	–0,1541	0,0744	4,29	0,0384
POL_ALL_1_CNT_V	01: low-0,5	0,1796	0,044	16,66	<,0001
POL_INS_REAL_LGTH_1_MY_V	01: low-2,5, MISSING	–0,3764	0,0804	21,89	<,0001
POL_INS_REAL_LGTH_1_MY_V	02:2,5–9,5	–0,0548	0,0829	0,44	0,5086
Intercept		–0,4334	0,0662	42,91	<,0001

Źródło: opracowanie własne.

Wszystkie zmienne oprócz ostatniej są statystycznie istotne na 5-procentowym poziomie istotności. Ze względu na występowanie dodatkowego poziomu grupowania, z których jeden jest statystycznie istotny, wskazana zmienna nie została usunięta z modelu.

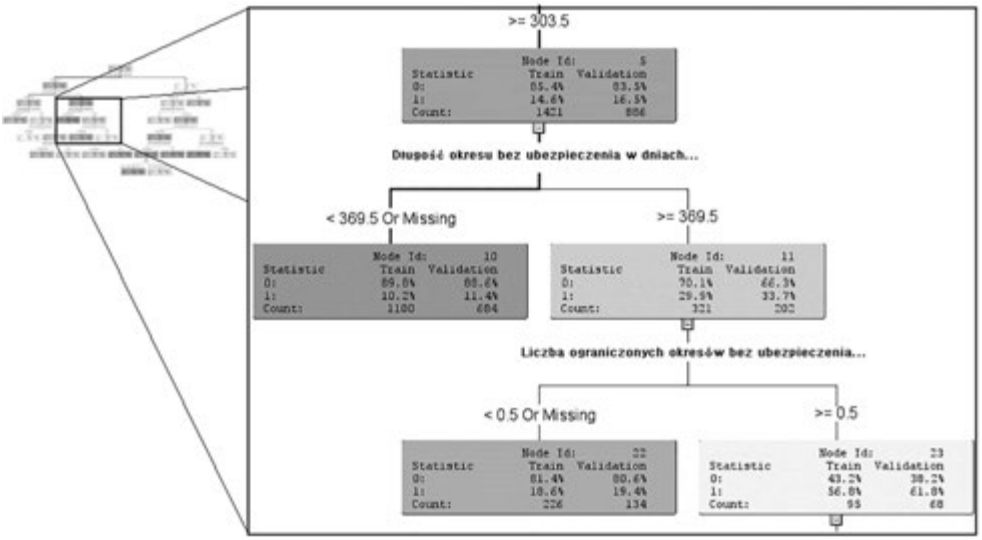
W przypadku modeli regresji logistycznej najczęściej interpretacji podlegają tzw. ilorazy szans (ang. *odds ratio*). W tabeli 6 podano ilorazy szans dla zmiennej POL_INS_REAL_LGTH_1_MY_V, oznaczającej długość rzeczywistego okresu ochrony ostatnio zawartej umowy ubezpieczenia dla właściciela (wartość wyrażona w pełnych miesiącach i na moment przeprowadzenia kontroli przez UFG). Ich interpretacja jest następująca:

1) skłonność do nieposiadania ochrony ubezpieczeniowej jest o ok. 55% mniejsza w przypadku, gdy faktyczna długość ochrony ostatnio zawartej umowy wyniosła do 2 miesięcy, niż w przypadku, gdy wartość ta wyniosła 10 miesięcy i więcej;
2) skłonność do nieposiadania ochrony ubezpieczeniowej jest o ok. 38% mniejsza w przypadku, gdy faktyczna długość ochrony ostatnio zawartej umowy wyniosła od 3 do 9 miesięcy, niż w przypadku, gdy wartość ta wyniosła 10 miesięcy i więcej.

Tabela 6. Ilorazy szans dla zmiennych w modelu regresji logistycznej – algorytm 1

Zmienna	Odniesienie	Iloraz szans
OPT_POL_INS_REAL_ LGTH_1_MY_V	01: low-2,5, MISSING vs 03:9,5-high	0,446
OPT_POL_INS_REAL_ LGTH_1_MY_V	02:2,5–9,5 vs 03:9,5-high	0,615

Źródło: opracowanie własne.



Rysunek 3. Model drzewa decyzyjnego – algorytm 1

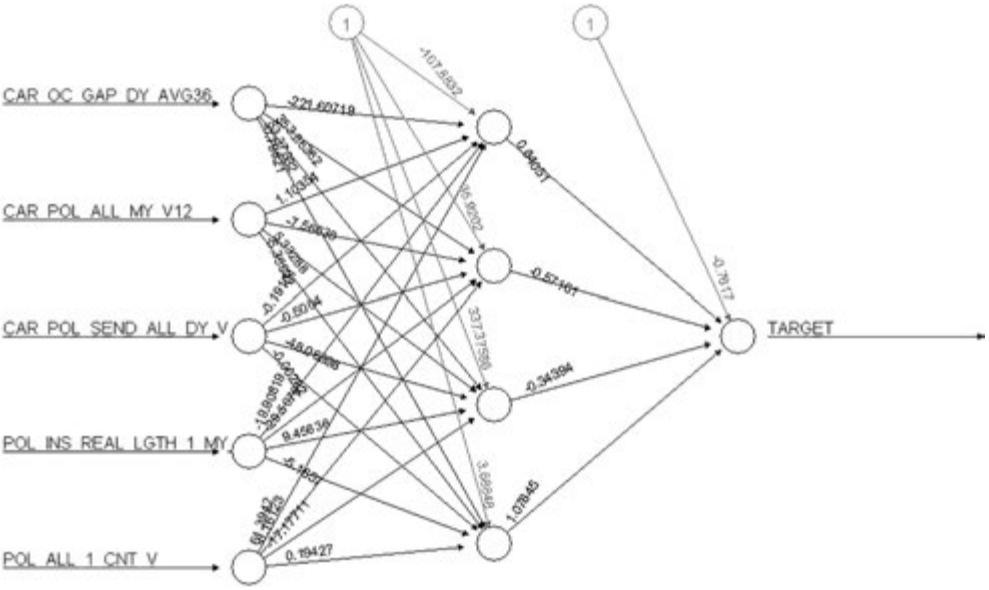
Źródło: opracowanie własne.

Na rysunku 3 przedstawiono jedno z drzew decyzyjnych stworzonych dla algorytmu 1. Drzewo przedstawia liczności przypadków z poszczególnych grup wraz z warunkami podziału. Wycinek drzewa przedstawia podział dokonany na podstawie dwóch zmiennych:

1) długości okresu bez ubezpieczenia dla pojazdu;
2) liczby ograniczonych okresów bez ubezpieczenia w historii pojazdu.

Interpretacja zależności dla wskazanego wycinka drzewa decyzyjnego jest następująca: w przypadku okresu bez ubezpieczenia dłuższego niż rok prawdopodobieństwo, że pojazd nie posiada ubezpieczenia, jest większe niż w przypadku, gdy okres ten wynosi około roku lub mniej (29,9% w porównaniu z 10,2%). Kolejna uzyskana zależność wskazuje, że w przypadku, gdy pojazdu nie dotyczyły ograniczone okresy bez ubezpieczenia, prawdopodobieństwo, że wykryty okres

bez ubezpieczenia występuje rzeczywiście, jest mniejsze niż w przypadku, gdy w historii pojazdu występował przynajmniej jeden taki okres.



Rysunek 4. Model sieci neuronowej – algorytm 1

Źródło: opracowanie własne.

Przedstawiona na rysunku 4 przykładowa sieć neuronowa ma warstwę wejściową składającą się z pięciu neuronów, odpowiadających liczbie wykorzystywanych zmiennych. Sygnały z neuronów wejściowych są przetwarzane w jednej warstwie ukrytej zbudowanej z czterech neuronów. W warstwie ukrytej oraz warstwie wyjściowej funkcję aktywacji stanowi funkcja logistyczna, z kolei wykorzystywaną funkcją błędu jest funkcja entropii. Warstwa wejściowa oraz ukryta mają dodatkowo neuron niezwiązany bezpośrednio z żadną ze zmiennych. Proces doboru wag między poszczególnymi neuronami odbywa się za pomocą metody zmodyfikowanej wstecznej propagacji błędów (ang. *resilient backpropagation*)[12].

W wyznaczonej sieci wartość zmiennej CAR_POL_ALL_MY_V12 z wagą 1,10354 jest umieszczana w pierwszym neuronie warstwy ukrytej. Wartość pierwszego neuronu zostaje wyznaczona za pomocą przekształcenia funkcją

[12] F. Günther, S. Fritsch, *neuralnet: Training of Neural Networks*, „The R Journal" 2010, vol. 2/1, June, s. 30–38.

aktywacji sumy wszystkich zmiennych wejściowych zmodyfikowanych o odpowiednią wagę. W ten sposób wyznaczona wartość pierwszego neuronu w warstwie ukrytej stanowi wartość wejściową, modyfikowaną o odpowiednią wagę, dla neuronu w warstwie wyjściowej. Jego wartość jest wyznaczana w taki sam sposób jak wartość pierwszego neuronu z warstwy ukrytej. W procesie uczenia sieci ta wartość neuronu warstwy wyjściowej jest porównywana z wartością rzeczywistą badanego przypadku.

4.2. Efektywność zastosowanego rozwiązania oraz monitorowanie skuteczności modeli statystycznych

Decyzje dotyczące wykorzystania modelu w działaniu produkcyjnym oparto na wskazaniach wybranych kryteriów efektywności dla zbiorów uczących oraz testowych. W przypadku modeli regresji logistycznej oraz drzew decyzyjnych pod uwagę brano również odpowiednio interpretację oszacowania parametrów oraz uzyskane kryteria podziału węzła.

W przypadku zarówno algorytmu 1, jak i 3 pod względem efektywności typowania nieubezpieczonych najlepiej sprawdzały się modele regresji logistycznej. W tabeli 7 zostały umieszczone informacje na temat wskaźników efektywności dla kilku oszacowanych modeli. Dalszej analizie poddano wskaźniki efektywności uzyskane na zbiorze testowym dla algorytmu 1 modelu regresji logistycznej.

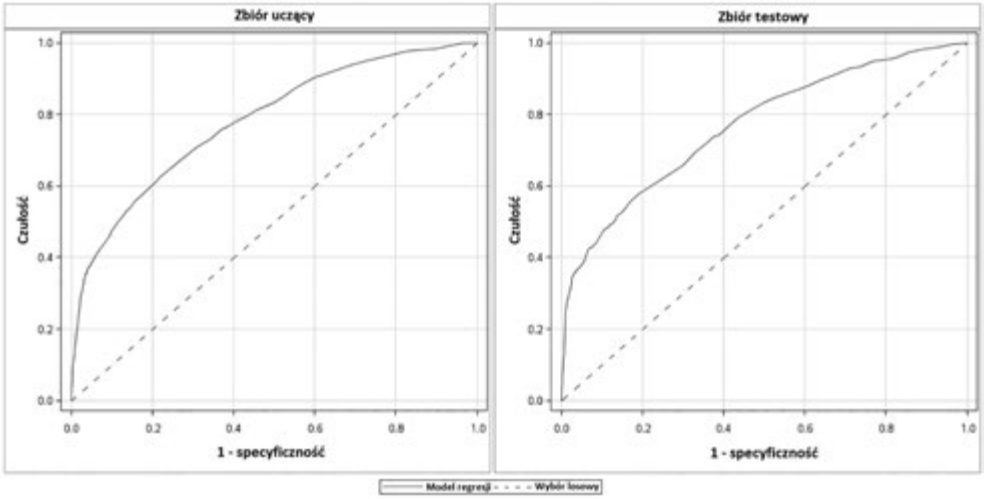
Tabela 7. Ocena efektywności modeli statystycznych na zbiorze testowym dla algorytmu 1 i algorytmu 3

Model predykcyjny	Algorytm	Prawdopodobieństwo błędnej klasyfikacji	Pole pod krzywą ROC	Wartość krzywej skumulowanego podbicia (5 percentyl)
Regresja logistyczna	1	0,21	0,77	2,9
Drzewa decyzyjne		0,28	0,72	2,51
Sieci neuronowe		0,23	0,75	2,78
Regresja logistyczna	3	0,25	0,8	2,1
Drzewa decyzyjne		0,33	0,75	2,01
Sieci neuronowe		0,31	0,71	1,98

Źródło: opracowanie własne.

Dla wskazanego algorytmu oraz modelu błędnie zostało przyporządkowanych 21% obserwacji. Wartość pola pod krzywą ROC wyniosła 0,77 (w przypadku

klasyfikatora, który losowo umieszcza obserwację w jednej z dwóch grup, wartość tego pola wynosi 0,5). Porównywalne wartości drugiego wskaźnika otrzymano dla zbioru uczącego, co świadczy o stabilności wyników. Krzywe ROC przedstawiono na rysunku 5.

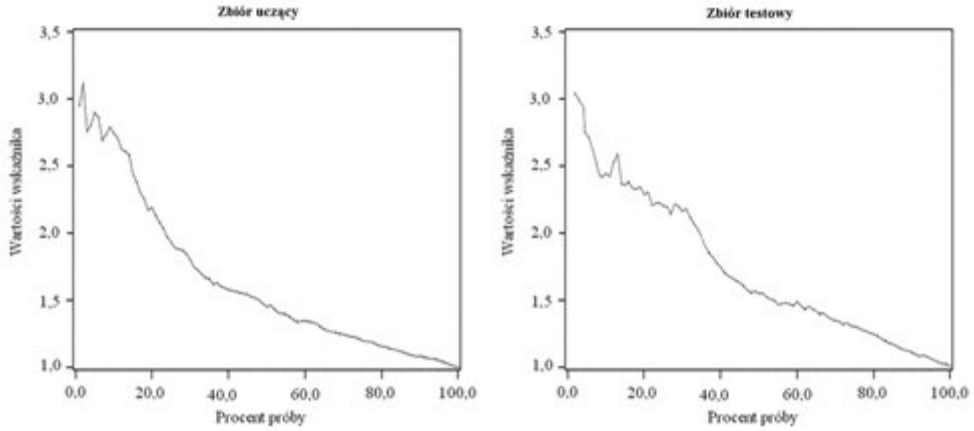


Rysunek 5. Krzywa ROC dla zbioru uczącego i testowego – algorytm 1 i model regresji logistycznej

Źródło: opracowanie własne.

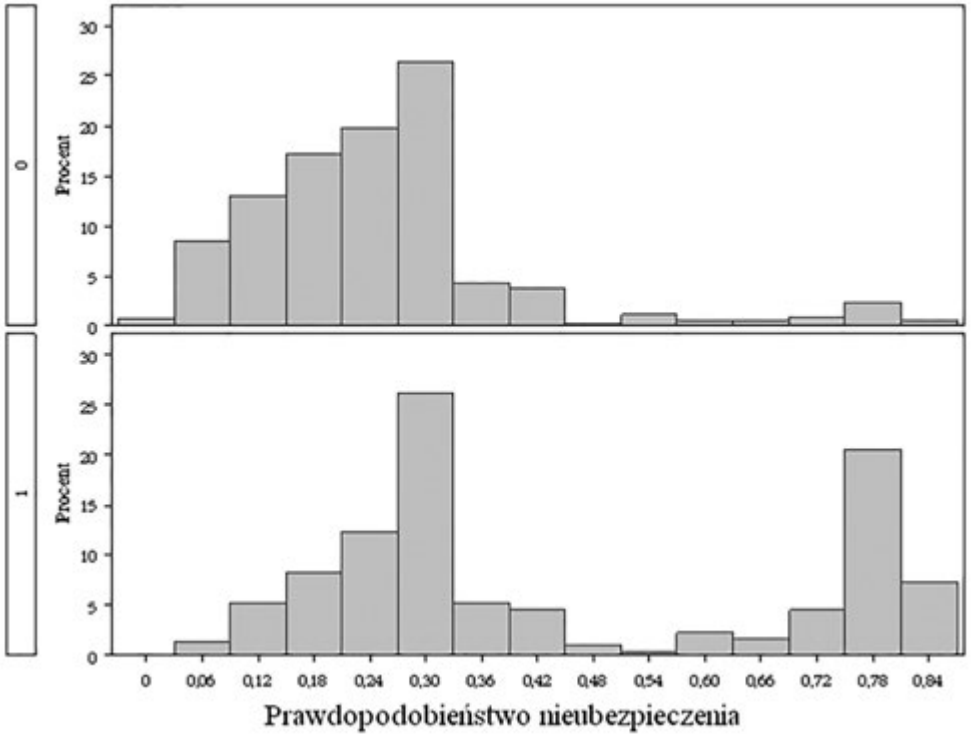
Wskaźnik skumulowanego „podbicia" dla 5 percentyla próby testowej wynosi 2,9, co oznacza, że wybierając pojazdy z listy uporządkowanej malejąco według prawdopodobieństwa braku ubezpieczenia, dla pierwszych 5% próby osiąga się prawie trzykrotnie wyższą skuteczność w identyfikacji nieubezpieczonych pojazdów i ich posiadaczy w stosunku do wyboru losowego. Krzywe skumulowanego „podbicia" przedstawiono na rysunku 6.

Rysunek 7 przedstawia rozkład prawdopodobieństwa nieubezpieczenia pojazdu w grupie przypadków zasadnych i niezasadnych wyznaczony na podstawie modelu regresji logistycznej na zbiorze testowym. Pojazdy z faktycznej grupy 0, przedstawione na pierwszym histogramie, są charakteryzowane przez wartości prawdopodobieństwa niższe od wartości charakteryzujących pojazdy z grupy 1, znajdujące się na drugim histogramie. Na podstawie rysunku można wstępnie potwierdzić właściwe działanie omawianego modelu.



Rysunek 6. Krzywa skumulowanego „podbicia" dla zbioru uczącego i testowego – algorytm 1 i model regresji logistycznej

Źródło: opracowanie własne.



Rysunek 7. Rozkład prawdopodobieństwa nieubezpieczenia w grupie przypadków niezasadnych i zasadnych – algorytm 1 i model regresji logistycznej

Źródło: opracowanie własne.

Monitorowanie skuteczności modelu jest możliwe na podstawie uzyskiwanych wyników. Niemniej ze względu na długi okres ustalania zasadności przypadków niezbędne jest wykorzystanie również innych wskaźników informujących o potrzebie zmiany lub wycofania modelu z użycia. Monitorowanie modelu polega na:

1) weryfikacji stabilności rozkładów wykorzystywanych zmiennych objaśniających między poszczególnymi cyklami procesu; weryfikacja dotyczy również rozkładu zmiennej objaśnianej z modelu (analiza podstawowych statystyk oraz histogramu między cyklami);
2) nadzorowaniu w zakresie zmian procesów zasilających bazy, a mających odzwierciedlenie w wyznaczanych zmiennych (zmiany: w uzupełnianiu braków danych; w trendach dotyczących terminowości zasilania przez zakłady ubezpieczeń; dotyczące zakresu gromadzonych danych).

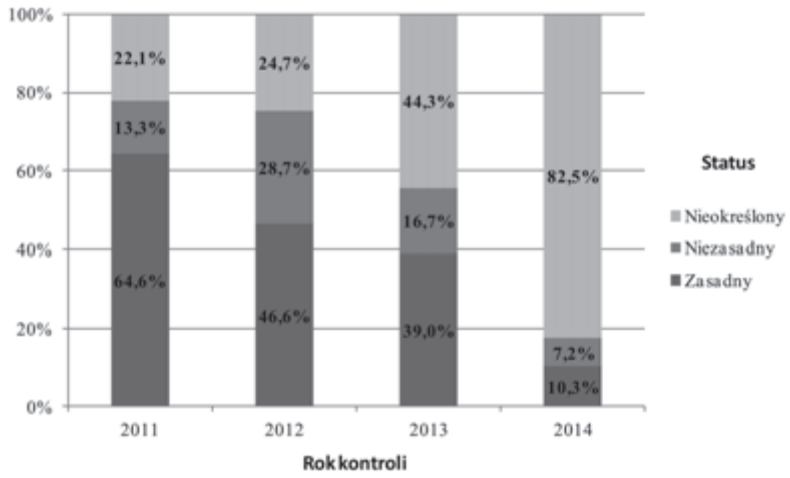
Zmiana modelu może również wynikać z potrzeby aktualizacji trendów występujących w odniesieniu do modelowanego zjawiska lub możliwości uzyskania większej liczby obserwacji dla procesów ponownego modelowania, w szczególności w wyniku otrzymanych informacji zwrotnych na temat przypadków wykorzystanych w poprzednich cyklach procesu.

4.3. Porównanie skuteczności kontroli UFG oraz podmiotów zewnętrznych

Efektywność kontroli własnych UFG spełnienia obowiązku posiadania ubezpieczenia OC p.p.m. jest wynikiem połączenia wykorzystywanych algorytmów typowania, modeli statystycznych oraz weryfikacji przypadków w zakładach ubezpieczeń. Od początku zautomatyzowanej kontroli do potencjalnie nieubezpieczonych posiadaczy pojazdów zostało wysłanych kilkadziesiąt tysięcy zawiadomień. W tym samym okresie kontrole były prowadzone przez podmioty zewnętrzne. Na rysunku 8 przedstawiono efektywność kontroli własnych UFG z wykorzystaniem systemu wykrywania nieubezpieczonych.

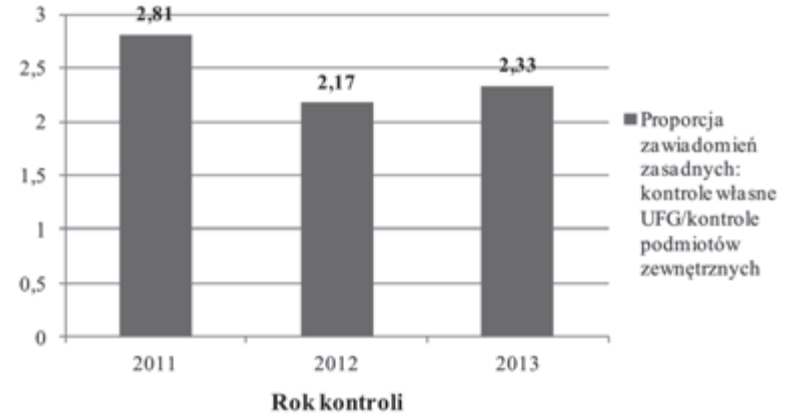
Niezależnie od źródła zawiadomienia można analizować w podziale na trzy grupy. Dla celów raportowych przyjęto, że w początkowej fazie sprawie opłatowej zostaje nadany status „nieokreślony". Wraz z upływem czasu status ten może zmienić się na „zasadny" lub „niezasadny". Jest to widoczne na rysunku 8, na którym w przypadku najnowszych spraw opłatowych status „nieokreślony" jest najczęstszy. Wraz z upływem czasu następuje stabilizacja proporcji spraw w poszczególnych statusach. Dodatkowo można zauważyć, że w początkowej fazie wzrost liczby przypadków niezasadnych jest większy niż w okresie „dojrzałości" portfela spraw opłatowych. Ze względu na charakterystykę procesu opłatowego

można również przyjąć, że zdecydowana większość przypadków ze statusem „nieokreślony" dla najwcześniejszych lat kontroli zmieni status na „zasadny". Skuteczność kontroli własnych UFG wyznaczona proporcją spraw zasadnych z wykorzystaniem systemu wykrywania nieubezpieczonych dla lat 2011–2014 wyniosła odpowiednio: 64,6%, 46,6%, 39%, 10,3% (dane z września 2014 r.). Szacuje się, że ostatecznie w procesie kontroli własnych siedem na dziesięć wytypowanych przypadków będzie zasadnych.



Rysunek 8. Efektywność kontroli własnych UFG w poszczególnych latach kontroli

Źródło: opracowanie własne.



Rysunek 9. Porównanie skuteczności kontroli własnych UFG z kontrolą przeprowadzaną przez podmioty zewnętrzne

Źródło: opracowanie własne.

Na rysunku 9 przedstawiono porównanie skuteczności zawiadomień z kontroli własnych UFG i kontroli przeprowadzanych przez podmioty zewnętrzne (ze względu na niewielką liczbę przypadków ze statusem innym niż „nieokreślony" pominięto 2014 r.). Wskazane współczynniki oznaczają, ile razy bardziej efektywna była kontrola przeprowadzona z wykorzystaniem systemu wykrywania nieubezpieczonych. Przykładowo, w 2013 r. proporcja spraw zasadnych w zawiadomieniach z kontroli własnych była ok. 2,33 razy większa niż w przypadku zawiadomień z podmiotów zewnętrznych. Niezależnie od tego, wyznaczony dla lat 2011–2013 stosunek spraw o statusie „nieokreślony" w przypadku kontroli własnych UFG do spraw o takim samym statusie w przypadku zawiadomień z podmiotów zewnętrznych wynosi 1,43. Wskazuje to na dużo większy potencjał zawiadomień z kontroli własnych UFG – część z nich przyjmie status „zasadny".

5. Podsumowanie

W niniejszej pracy przedstawiono rozwiązanie zastosowane w Ośrodku Informacji Ubezpieczeniowego Funduszu Gwarancyjnego, mające na celu kontrolę spełnienia obowiązku posiadania ubezpieczenia odpowiedzialności cywilnej posiadaczy pojazdów mechanicznych. O kompleksowości zagadnienia świadczą wykorzystywane w celu ustalenia prawdopodobieństwa nieubezpieczenia statystyczne systemy uczące się pod nadzorem, ale również udoskonalane algorytmy typowania, algorytmy łączące pojazdy w klastry czy zaangażowanie zakładów ubezpieczeń w proces weryfikacji przypadków.

Początkowo system wykrywania nieubezpieczonych opierał się wyłącznie na algorytmie wyznaczającym okres bez ubezpieczenia *post factum*. Dzięki dodatkowym algorytmom kontrole przeprowadzane przez UFG można porównać z tymi wykonywanymi przez podmioty zewnętrzne bezpośrednio podczas kontroli drogowych lub w trakcie czynności urzędowych. Dzięki ich wdrożeniu system zyskał charakter prewencyjny, wskazując tym samym główny cel działań prowadzonych przez UFG. Celem tym jest ograniczenie liczby nieubezpieczonych posiadaczy pojazdów uczestniczących w ruchu drogowym. Korzyści, które zostaną osiągnięte dzięki realizacji tego celu, mają bezpośrednie przełożenie na rynek ubezpieczeń komunikacyjnych i będą powodowały[13]:

[13] Por. W. Bijak, K. Hrycko, S. Garstka, *Automatyzacja prowadzonych przez UFG kontroli spełnienia obowiązku zawarcia umowy ubezpieczenia OC posiadaczy pojazdów mechanicznych*, „Prawo Asekuracyjne" 2013, nr 3, s. 74.

1) zmniejszenie liczby zdarzeń, za które odpowiedzialność finansową (z możliwością późniejszego regresu) ponosi UFG;
2) zwiększenie przychodów zakładów ubezpieczeń z tytułu dodatkowego przypisu składki;
3) ograniczenie wielu prywatnych i społecznych konsekwencji, które wynikają z obowiązku poniesienia kosztów następstw wypadku w okresie bez ubezpieczenia.

Należy zwrócić uwagę na fakt, że każdy nieubezpieczony to dodatkowy koszt dla pozostałych ubezpieczających się użytkowników pojazdów. Szacuje się, że ze względu na opisywane zjawisko w Wielkiej Brytanii umowa ubezpieczenia jest droższa o około 30 GBP[14].

W systemie wykrywania nieubezpieczonych nie można nie docenić wartości poszczególnych składowych zapewniających odpowiednią jakość danych zarówno wejściowych, jak i wyjściowych. W zapewnieniu jakości danych wyjściowych szczególną rolę odgrywają statystyczne systemy uczące się pod nadzorem, pozwalające na typowanie przypadków z najwyższym prawdopodobieństwem zasadności. Porównanie dotyczące wyboru losowego z typowanej grupy potencjalnie nieubezpieczonych, jak również zawiadomień z podmiotów zewnętrznych wskazało na wysoką skuteczność wykorzystywanych modeli predykcyjnych oraz zastosowanego podejścia.

System wykrywania nieubezpieczonych jest nieustannie rozwijany. Nie tylko dotyczy to weryfikacji obecnych oraz budowy nowych modeli statystycznych dla algorytmów typowania, ale obejmuje również wykorzystanie automatycznych procesów systemu w przypadku zawiadomień o braku ubezpieczenia, które trafiają do UFG z różnych źródeł. W związku z tym w przyszłości podobne działania, jak prowadzone obecnie w systemie wykrywania nieubezpieczonych, będą realizowane w przypadku zawiadomień z podmiotów zewnętrznych czy zawiadomień anonimowych. Dotyczy to weryfikacji w bazie OI oraz weryfikacji przeprowadzanej w zakładach ubezpieczeń, a w miarę rozwoju systemu wykorzystania modelu lub modeli predykcyjnych ustalających zasadność tych przypadków. Ostatecznym efektem rozwoju kontroli spełnienia obowiązku posiadania ubezpieczenia OC p.p.m. będzie obsługa zdecydowanej większości zawiadomień w sposób automatyczny wraz z kierowaniem przypadków na ustalone ścieżki do dalszego procedowania przez pracowników UFG.

[14] http://www.mib.org.uk/NR/rdonlyres/93DB4017-FD8B-45DF-889A-8DCB9905962F/0/MIB reportFINAL.pdf (dostęp: 12.09.2014).

Bibliografia

Bijak W., Hrycko K., Garstka S., *Automatyzacja prowadzonych przez UFG kontroli spełnienia obowiązku zawarcia umowy ubezpieczenia OC posiadaczy pojazdów mechanicznych*, „Prawo Asekuracyjne" 2013, nr 3, s. 71–82.

Bijak W., Hrycko K., Pietrowski Ł., *Sposób na nieubezpieczonych*, „Miesięcznik Ubezpieczeniowy" 2012, kwiecień, s. 15–18.

Ćwik J., Mielniczuk J., *Statystyczne systemy uczące się. Ćwiczenia w oparciu o pakiet R*, Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2009.

Günther F., Fritsch S., *neuralnet: Training of Neural Networks*, „The R Journal" 2010, vol. 2/1, June, s. 30–38.

Hastie T., Tibshirani R., Friedman J., *The Elements of Statisical Learning: Data Mining, Inference, and Prediction*, Springer, New York 2009.

Koronacki J., Ćwik J., *Statystyczne systemy uczące się*, Akademicka Oficyna Wydawnicza EXIT, Warszawa 2008.

Larose D.T., *Data Mining Methods and Models*, John Wiley & Sons, Hoboken 2006.

Skorzybut M., Krzyśko M., Górecki T., Wołyński W., *Systemy uczące się. Rozpoznawanie wzorców, analiza skupień i redukcja wymiarowości*, Wydawnictwa Naukowo-Techniczne, Warszawa 2008.

Tadeusiewicz R., *Sieci neuronowe*, Akademicka Oficyna Wydawnicza, Warszawa 1993.

Ustawa z dnia 22 maja 2003 r. o ubezpieczeniach obowiązkowych, Ubezpieczeniowym Funduszu Gwarancyjnym i Polskim Biurze Ubezpieczycieli Komunikacyjnych (Dz. U. z 2011 r. Nr 205, poz. 1210).

Zaawansowane metody analiz statystycznych, red. E. Frątczak, Oficyna Wydawnicza SGH, Warszawa 2012.

Źródła sieciowe

http://www.mib.org.uk/NR/rdonlyres/93DB4017-FD8B-45DF-889A-8DCB9905962F/0/MIBreportFINAL.pdf (dostęp: 12.09.2014).

* * *

Methods and statistical models used to identify uninsured car owners

Summary

The paper presents the approach used by the Polish Insurance Guarantee Fund to tackle the phenomenon of uninsured car owners. The process covers different areas concerning:

1. algorithms which identify discontinuity in MTPL coverage,
2. the usage of statistical modelling called supervised learning, including such models as: generalised linear models, decision trees and neural networks,
3. the cooperation with insurers to identify the uninsured.

The paper presents the results obtained in exemplary statistical models, as well as the achieved accuracy of prediction. The publication presents further evolution of the developed system.

Keywords: MTPL insurance coverage control, automatic identification of the uninsured, supervised learning