

**doskonalenie działań  
administracji podatkowej**  
z wykorzystaniem drążenia danych

## Wstęp

Zarówno federalne jak i stanowe agencje podatkowe w Stanach Zjednoczonych zostały powołane do kontroli i maksymalizacji ściągalności podatków. Celem tego opracowania jest pokazanie, jak techniki drążenia danych mogą pomóc agencjom skarbowym w osiągnięciu celów, jakimi są efektywna kontrola i wysoka ściągalność podatków. Prezentację problematyki rozpoczyna ogólna informacja, czym są techniki drążenia danych, następnie przedstawione zostaną szczegóły dotyczące systemu wdrożonego w Audit Division. Wdrożenie przygotowane zostało przez Texas Comptroller of Public Accounts (CPA), oraz partnera SPSS Inc. – Elite Analytics, LLC, firmę konsultingową zajmującą się wdrażaniem systemów data mining. Opracowanie kończy się przeglądem dodatkowych zastosowań technik drążenia danych w obszarze ściągania podatków.

Agencje podatkowe stosują kontrolę, by zagwarantować wysoką ściągalność podatków oraz utrzymać stały strumień dochodów. Kontrola pośrednio podwyższa dobrowolne płacenie podatków oraz bezpośrednio generuje dodatkową ich ściągalność, co pomaga zmniejszyć obszar szarej strefy. Dlatego kontrola jest istotnym elementem egzekwowania prawa podatkowego i zapewnia stabilność dochodów państwa.

Efektywne zarządzanie agencjami kontrolującymi podatników wymaga jednak podejmowania wielu decyzji. Z punktu widzenia zagadnień omawianych w tym opracowaniu warto wymienić takie, często zadawane przez zarządzających, pytania:

- Czy lepiej jest zbudować jeden schemat doboru podmiotów do kontroli, czy też efektywniejsze będą kombinacje wielu różnych?
- Czy lepiej schematy te budować w oparciu o identyfikowaną wysokość placonych podatków, czy typ podatnika?
- Jak powinno się rozdzielić zasoby do kontroli podatników płacących różne typy podatków?

Skupienie się na kontroli podatników płacących różne typy podatków może skutkować zwiększeniem dochodów podatkowych. Inna może być też częstotliwość występowania skłonności do unikania płacenia podatków w przypadku każdego z nich. Samo przeprowadzenie kontroli powinno być rozumiane jako wieloetapowy proces: od wytypowania podmiotu do kontroli, poprzez przeprowadzenie kontroli, stwierdzenie wystąpienia nieprawidłowości, podjęcie negocjacji, po ściągnięcie należności czy, niekiedy, prawną windykację należnego podatku. Na każdym z tych etapów można podejmować decyzje zwiększające lub zmniejszające efektywność całego procesu kontrolnego.

Można korzystać z różnych metod doboru podmiotów do kontroli, od losowej selekcji, poprzez bardziej złożony system typowania oparty na regułach wychwytyjących interesujące przypadki, aż do zaawansowanych technik statystycznych i drążenia danych. Możemy, na przykład, zdecydować się na wybór strategii bazującej na podziale podmiotów na segmenty (np. płacących różne typy podatków), a następnie zastosować do każdego z nich odrębne reguły typujące podmioty do kontroli. Do tej pory w stanie Teksas wyodrębnia się górny segment podmiotów płacących łącznie 65% wartości uzyskanego podatku od sprzedaży. Nazywany jest on „Priority One” i stosuje się do niego regułę mówiącą, że należy skontrolować z tego segmentu każdego przynajmniej raz w przeciągu czterech lat.

W Teksasie kontroluje się również inną grupę podmiotów, nazywaną „Prior Productive”. Są to wszystkie te podmioty, którym podczas ostatniej kontroli wykazano i wyegzekwowano niezapłacone podatki o wartości przewyższającej \$10 000.

Tak jak w wielu stanach, populacja płacących podatki w Teksasie rośnie jednostajnie od ostatniej dekady, bez proporcjonalnego wzrostu nakładów na kontrolę. W rezultacie Teksas i kilka innych stanów, a także agencje podatkowe w Wielkiej Brytanii i Australii, postawiły na techniki drążenia danych pomocne w odnajdowaniu osób zalegających z płacaniem podatków oraz w podejmowaniu bardziej efektywnych decyzji. Drążenie danych wymaga systemów zbierania i przechowywania danych, integrujących zewnętrzne i wewnętrzne źródła danych. Umożliwia to zastosowanie wielu typów analiz, od analiz trendów do odpowiedzi na pytania:

- Jak powinniśmy rozłożyć nakłady na działania kontrolne?
- Jak wyselekcjonować podatników, którzy powinni zostać poddani kontroli?
- Jaki jest przewidywany wzrost dochodu z podatków w wyniki poszczególnych typów kontroli?

## Dlaczego data mining?

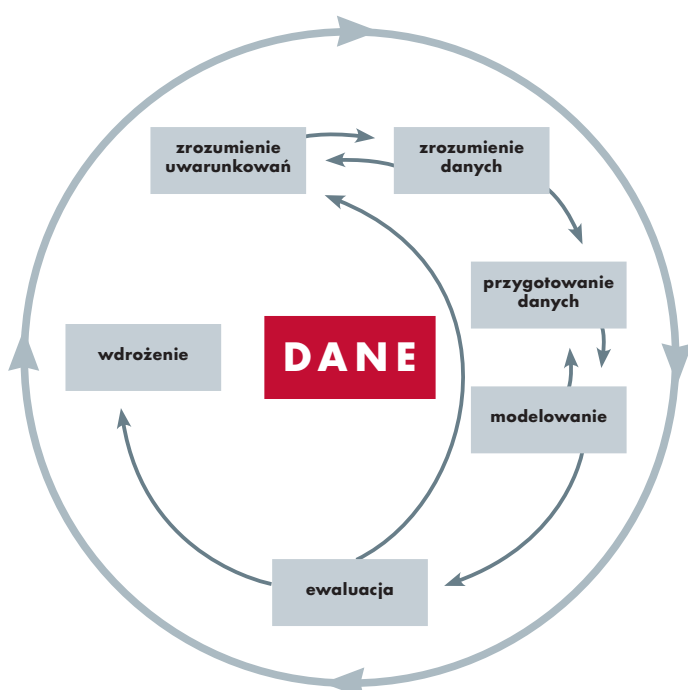
Instytucje kontrolne mają dostęp do ogromnej ilości danych na temat podatników i rzeczywiście większość z nich czerpie informacje z tych źródeł, aby wspomóc swoje działania. Przykładowo, na podstawie dostępnych informacji poszukuje się, których podatników można opisać przy pomocy określonych przez ekspertów specjalnych profili. Mogą być one oparte o pojedynczy wymiar, cechę (np. typ działalności gospodarczej) lub ich kombinację (np. typ przedsiębiorstw handlowych posiadających określony stosunek wartości zapłaconego podatku do wartości sprzedaży). W tym miejscu należy podkreślić, że techniki drążenia danych umożliwiają to samo, ale na znacznie większą skalę. Używając technik drążenia danych instytucje kontrolne mogą analizować dane o setkach tysięcy podatników, aby zidentyfikować ich wspólne cechy, a potem konstruować, na tej podstawie, profile reprezentujące różne typy zachowań podatników. Dzięki technikom drążenia danych, instytucje te mogą znaleźć podobieństwa pomiędzy podmiotami charakteryzującymi się dużym zwrotem podatku po kontrolach i dzięki temu skupić się na sprawdzaniu innych podmiotów posiadających podobne atrybuty. Drążenie danych umożliwia zrozumienie danych, a także analizę i prognozę zachowań osób unikających płacenia podatków.

## Co to jest drążenie danych?

Drążenie danych to automatyczna lub pół-automatyczna eksploracja i analiza dużych zbiorów złożonych danych, w celu odkrycia złożonych wzorców i reguł. Organizacje mogą używać tego typu analiz, aby przeciwdziałać oszustwom podatkowym i zidentyfikować zachowania wskazujące na działania zmierzające do uniknięcia płacenia podatków. Drążenie danych umożliwia eksplorację i analizę danych bez posiadania gotowych hipotez. To ważna konstatacja, gdyż tradycyjna analiza statystyczna oparta jest na testowaniu wcześniej wyrażonych hipotez. Pomimo bardzo mocnego eksploracyjnego charakteru analiz data mining, konieczne jest stosowanie przez organizacje usystematyzowanego podejścia do procesu realizacji takich projektów.

Metodologia CRISP-DM (The CRoss-Industry Standard Process for Data Mining) – zob. rysunek 1.) została utworzona w 1996 r. przez konsorcjum konsultantów z zakresu drążenia danych i specjalistów od technologii z SPSS, Daimler-Benz (teraz Daimler-Chrysler) i NCR. Twórcy CRISP-DM, odnosząc się do własnych doświadczeń, rozwinęli sześciofazowy proces umożliwiający ujęcie celów biznesowych i wiedzy organizacyjnej przy realizacji projektów. Metodologia CRISP-DM, postrzegana jest obecnie jako standard przy realizacjach projektów data mining.

### Sześć faz CRISP-DM



Rysunek 1.  
CRISP-DM

### Zrozumienie uwarunkowań

Przeprowadzenie pierwszej fazy zapewnia, iż wszyscy uczestnicy rozumieją cele projektu z perspektywy biznesowej i własnej organizacji. Cele te są potem formułowane w trakcie definiowania problemu drążenia danych i przygotowania planu projektu. W instytucji kontrolnej faza ta powinna oznaczać zdefiniowanie i zrozumienie procesu zarządzania kontrolą, jej funkcje, odpowiedź na pytanie: jakie role spełniają jej poszczególne etapy oraz jakie dane są gromadzone. Ważne jest by jasno zostało określone na czym polega podniesienie efektywności w tego typu instytucjach. Informacje te powinny zostać zawarte w definicji problemu drążenia danych i planie projektu.

### Zrozumienie danych

Druga faza zakłada analizę źródeł, ich jakości i charakterystyk danych. Ta początkowa eksploracja umożliwia wgląd i pomaga skupić się na projekcie. Rezultaty jej pomagają zrozumieć dane, które będą wykorzystywane w budowaniu modeli. Ta faza może być czasochłonna dla tych instytucji, które mają wiele źródeł danych, ale jest to wyjątkowo ważny moment w projekcie.

### Przygotowanie danych

Następna faza polega na umieszczeniu danych w takiej formie, która jest odpowiednia dla budowania modeli. Analitycy na podstawie wcześniej wyznaczonych celów określają, które dane i algorytmy drążenia danych użyć. W tej fazie trzeba także rozwiązać problemy nie zidentyfikowane we wcześniejszej fazie, jak np. występowanie braków danych.

### Modelowanie

Ta faza sprowadza się do budowania algorytmów drążenia danych, pozwalających wydobywać informacje z danych. Istnieje wiele różnych technik drążenia danych, które są odpowiednie i umożliwią pozyskanie potrzebnych informacji. Agencje podatkowe mogą na przykład używać klasyfikacyjnych lub regresyjnych modeli, żeby odkryć cechy kontroli podatkowych, które były efektywniejsze. Każda technika potrzebuje określonego rodzaju danych, co może wymagać powrotu do fazy przygotowywania danych. Efektem tej fazy jest model lub zestaw modeli wykorzystujący w odpowiedni sposób wiedzę odkrytą dzięki analizom.

### Ewaluacja

Celem tej fazy jest ocena jakości modelu lub modeli. Zastosowanie algorytmów drążenia danych może skutkować ujawnieniem nieograniczonej liczby wzorców, jednak wiele z nich może nic nie wnosić do rozwiązania problemu. W tej fazie trzeba wyodrębnić, które modele są użyteczne do uzyskania odpowiedzi na postawione pytania. W przypadku problemu związanego z odpowiednim doбором podatników do działań kontrolnych, model predykcyjny powinno testować się na danych historycznych. Wtedy wyniki modelu można porównać z wynikami przeprowadzonych już kontroli.

### Wdrożenie

W tej fazie wyniki określone przez „pracujące” modele włączone zostają do realnego procesu podejmowania decyzji. W zależności od rezultatów, zbudowany system wspomagający proces decyzyjny wymagać może tylko niewielkich modyfikacji lub wprowadzenia większych zmian i reorganizacji w systemie. Faza wdrożenia zezwala również na przejście ponownie procesu w celu ulepszenia i dostosowania modeli. Czynnikiem wymuszającym powtórzenie omawianego procesu są, na przykład, częste zmiany prawa podatkowego.

Sześć powyższych faz jest obecne w każdym procesie drążenia danych. Mimo iż każda faza jest ważna to ich kolejność nie jest sztywna. Proces drążenia może wymagać poruszania się tam i z powrotem między fazami. Następną fazę lub zadanie zależy od wyników poprzedniej fazy. Wewnętrzne strzałki na rysunku 1. wskazują najważniejsze i najczęściej występujące zależności pomiędzy fazami. Zewnętrzne koło symbolizuje cykliczną naturę procesu drążenia danych, gdyż praca wykonana podczas projektu pozwala często na stawianie kolejnych lepiej sprecyzowanych problemów. Ogólnie, doświadczenie zdobyte przy realizacji wcześniejszych projektów, procentuje przy pracy nad następnymi. Do 90% czasu realizacji każdego projektu drążenia danych pochłaniają początkowe fazy poprzedzające proces modelowania danych. Przede wszystkim od wyników prac w tych właśnie fazach zależy sukces całego projektu. W samym procesie budowy modeli analitycy zazwyczaj wykorzystują gotowe zestawy technik i narzędzi.

Techniki data mining pochodzą z różnych subdyscyplin. Wśród nich wykorzystuje się algorytmy sztucznej inteligencji, algorytmy statystyczne, ewaluacyjne czy modele pozwalające na odnajdywanie i wizualizację wzorców i reguł w danych. Szczegółowa dyskusja na temat różnorodnych aspektów wykorzystania technik drążenia danych wykracza poza ramy tego opracowania. W kolejnej części zaprezentujemy kilka rozwiązań, które zostały wdrożone w instytucjach zajmujących się kontrolą podatków.

## Studium przypadku: strategia typowania podmiotów do kontroli w stanie Teksas

Jak wcześniej wspomniano, istnieje wiele metod doboru podmiotów do kontroli, a rezultaty każdej z nich różnią się poziomem efektywności. Zależność pomiędzy efektywnością a strategią wyboru podmiotów do kontroli staje się wymierna, gdy przyjmiemy za wskaźnik efektywności kwotę odzyskanych podatków na godzinę przeprowadzanego audytu. W tym rozdziale zostanie przedstawiony Audit Select – system scoringowy wdrożony przez Elite Analytics w Audit Division of the Texas Comptroller of Public Accounts. Zostanie on również porównany z tradycyjną metodą wyboru podmiotów do kontroli stosowaną przez tę instytucję.

Clementine to rozwiązanie do analiz data mining, które zostało wybrane i użyte do wdrożenia nowej strategii doboru podmiotów do kontroli. Według analityka Elite: „W Clementine znaleźliśmy nie tylko różnorodne możliwości przygotowania danych, tworzenia i zastosowania modeli, które trudno znaleźć w innych produktach, ale również sama praca z interfejsem Clementine dostosowanym do metodologii CRISP-DM była i jest bardzo intuicyjna”.

## Modele predykcyjne

Modele predykcyjne są bodaj najczęściej spotykanym przykładem wykorzystania analiz data mining. Jak sugeruje nazwa, modele predykcyjne wykorzystywane są do przewidywania występowania danych wartości, zdarzeń czy procesów. W przypadku agencji kontrolujących podatków, przewiduje się, u których podmiotów najprawdopodobniej wykaże się i wyegzekwuje względnie największą wartość niezapłaconego podatku.

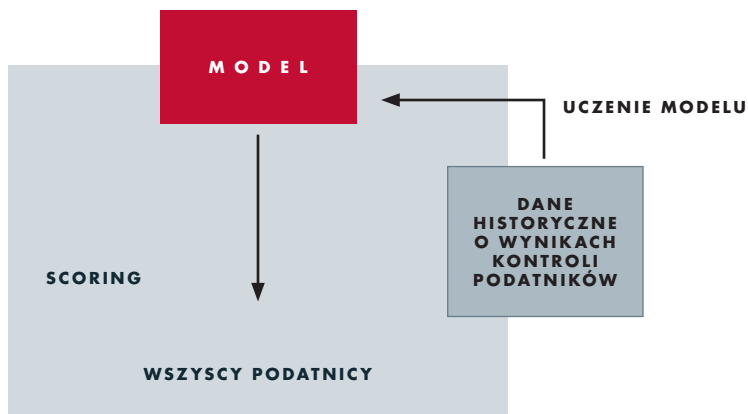
Modele predykcyjne zazwyczaj tworzą scoring, numeryczną skalę, która implikuje prawdopodobne wyniki audytu. Wysoka wartość scoringu może wskazywać, że korekta podatku wynikająca z kontroli będzie wysoka lub ponadprzeciętna, podczas gdy niska wartość scoringu sugerować będzie niskie prawdopodobieństwo wysokiej korekty podatkowej. Przykładem jest Indeks DIF (Discriminant Index Function) stworzony przez IRS do identyfikacji deklaracji podatkowych, w których z wysokim prawdopodobieństwem nie ujawniono wszystkich dochodów. Rezultatem używania tego indeksu jest znacznie wyższa efektywność kontroli w porównaniu do przeprowadzanych całkowicie losowo.

Model scoringowy stał się bardzo popularny głównie dzięki możliwości uwzględnienia w modelu setek zmiennych dla bardzo dużej liczby rekordów. Poza wieloma innymi zastosowaniami, organizacje używają modeli scoringowych do:

- szacowania ryzyka kredytowego,
- identyfikowania oszustw dokonywanych przy pomocy kart kredytowych,
- określania potencjalnych odpowiedzi na wysyłki marketingu bezpośredniego,
- porządkowania ofert w kolejności wyznaczonej prawdopodobieństwem zakupu.

### Zastosowanie modeli do predykcji wyników kontroli podatkowych

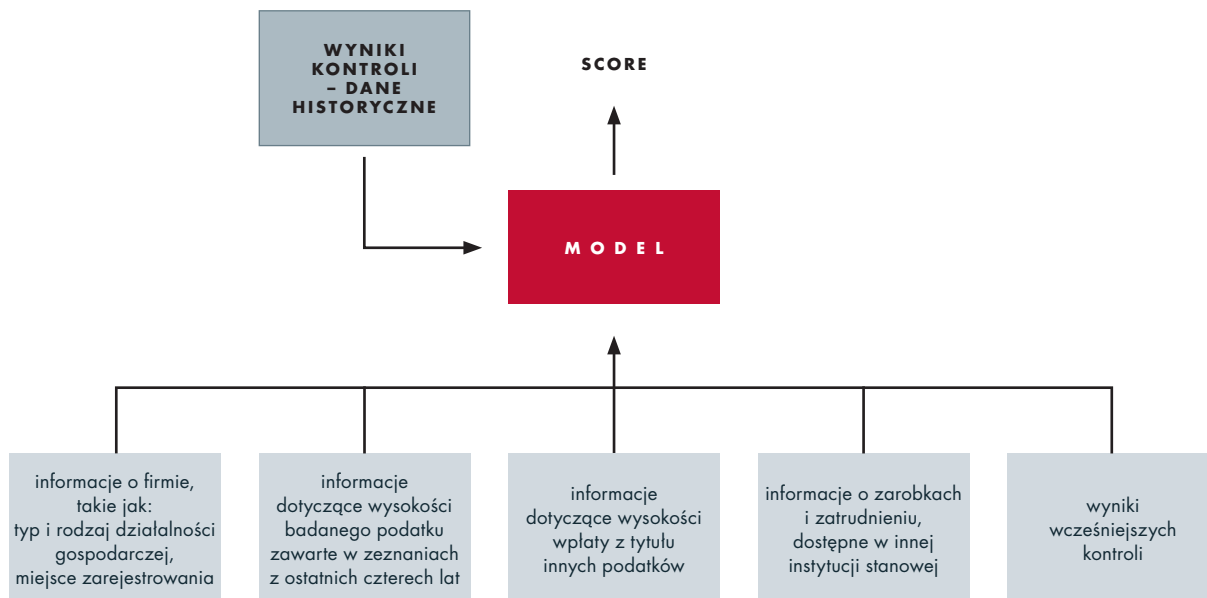
Baza danych (Advanced Database System, ADS) jest hurtownią danych dedykowaną do wspomagania systemów analitycznych, takich jak Audit Select, w którym zastosowany został model predykcyjny do identyfikacji podmiotów do kontroli. Rysunek 2. obrazuje proces budowania i używania modeli predykcyjnych do wyboru grupy przeznaczonej do kontroli.



**Rysunek 2.**  
Trening i wykorzystanie modelu

Pierwszy krok to „uczenie” modelu przy użyciu danych historycznych o podatnikach, dla których wyniki kontroli są znane, czyli zależności pomiędzy cechami podatnika, a wynikiem kontroli. Modele scoringowy Audit Select używa następujących pięciu źródeł danych aby stworzyć profil podatnika:

- informacje o firmie, takie jak: typ i rodzaj działalności gospodarczej, miejsce zarejestrowania,
- informacje dotyczące wysokości podatku, który jest badany zawarte w zeznaniach z ostatnich czterech lat,
- informacje dotyczące wysokości wpłat z tytułu innych podatków,
- informacje o zarobkach i zatrudnieniu, dostępne w innej instytucji stanowej,
- wyniki wcześniejszych kontroli.



**Rysunek 3.**  
Źródła danych w systemie Audit Select

„Nauczony” model może być zastosowany do całej populacji podatników. Do nowych podmiotów stosuje on reguły, jakich nauczył się podczas analizy wyników wcześniejszych kontroli (zob. rysunek 4.). Następnie kontrolerzy używają wartości scoringu do określenia które instytucje będą podlegać kontroli.

Modele budowane są w oparciu o dane będące wynikiem wielu przekształceń i transformacji zamieniających surowe dane w bardziej syntetyczne wskaźniki. Dla przykładu, gdy dobrym wskaźnikiem okazuje się wartość wysokości odpisów (sama w sobie lub w odniesieniu do innych wartości, takich jak wysokość przychodów), to dobre efekty może wykazać porównanie jego ocen wśród podobnych podmiotów. Modele mogą później bazując na takich wewnątrz grupowych porównaniach, posłużyć do budowy kolejnych wskaźników.

Na rysunku 4. została przedstawiona pierwsza tabela, która prezentuje zestaw danych historycznych przeznaczonych do uczenia modelu. Druga tabela pokazuje wartości scoringu wyliczonego przez model dla obecnej populacji podatników.

#### DANE HISTORYCZNE UCZĄCE MODEL

obrót	klasyfikacja SIC	wynagrodzenia	wpływy podatkowe	odzyskany podatek
\$ 21 110 288	23	\$ 34 456 345	\$ 988 945	\$ 100 202
\$ 34 234 334	43	\$ 11 476 544	\$ 2 545 251	\$ 434 323
\$ 9 874 556	23	\$ 45 443 343	\$ 4 534 521	\$ 0
\$ 33 421 655	56	\$ 45 433	\$ 4 354 353	\$ 454 352
\$ 254 667 678	55	\$ 445 453	\$ 43 657 176	\$ 82 834

#### PRZEWIDYWANE WYNIKI KONTROLI

obrót	klasyfikacja SIC	wynagrodzenia	wpływy podatkowe	przewidywany score
\$ 424 454 762	43	\$ 5 454 362	\$ 22 571 243	760
\$ 44 572 462	32	\$ 45 445 623	\$ 45 653 235	450

### Rezultaty pracy systemu scoringowego Audit Select

Najbardziej efektywnym sposobem oceny systemu scoringowego jest porównanie wyliczonych przez niego dla każdego z podatników wartości implikujących prawdopodobne wyniki audytu z rzeczywistymi wynikami kontroli. W przypadku agencji kontrolujących podatników porównanie polega na zmierzeniu zależności pomiędzy faktyczną korektą podatku, a wskaźnikiem wpływów obliczonym przez model.

Zgodnie z oczekiwaniami średnia korekta podatku rośnie proporcjonalnie do wzrostu wartości scoringu. Na przykład kontrola dokonana na podatnikach, dla których wartość scoringu wynosiła mniej niż 100 punktów wykazała, iż ich korekta podatkowa wynosiła średnio \$3300. Inaczej niż u podatników dla których wartość scoringu wynosiła ponad 900 punktów, dla nich korekta podatkowa wyniosła przeciętnie \$78000.

Pomimo tego, że liczba podmiotów posiadających scoring na poziomie ponad 900 punktów jest znacznie mniejsza od liczby, którym model przyporządkował wartości poniżej 100, to, jak wynikało z dodatkowych analiz, tylko 57% badanych z wyższymi ocenami było przynajmniej raz kontrolowanych. Jak pokazuje to porównanie, ten segment podatników powinien być poddawany szczegółowej kontroli.

**Rysunek 4.**  
Wyniki przeprowadzonych testów

## Porównanie z innymi metodami doboru podmiotów do kontroli

Jak wykazano powyżej techniki drążenia danych i modele predycyjne użyte w systemie scoringowym Audit Select wniosły nową jakość do dotychczas istniejącego procesu doboru podatników do kontroli. Jakkolwiek tym, co jest najbardziej adekwatną miarą dla nowych technologii, nie jest poziom ich technologicznej złożoności czy ukryta w nich naukowa wiedza, ale stopień, w jakim podnoszą one efektywność dotychczasowego procesu kontroli. Takie porównanie jest w szczególności ważne, gdy dotychczas stosowało się mniej złożone, ale dobrze przetestowane metody doboru podmiotów do kontroli.

## Inne strategie kontroli podatkowej

Przed przedstawieniem oceny przedstawianego tutaj systemu scoringowego Audit Select, pokażemy w skrócie, od lat stosowane w Teksasie sposoby doboru podmiotów do kontroli.

### Priority One (pierwsza grupa podatników)

W stanie Teksas najwięksi podatnicy, dostarczający 65% wszystkich wpływów z tytułu podatków zaliczani są do odrębnej grupy podatników (tzw. Priority One). Do tych podatników stosowano regułę nakazującą kontrolowanie ich przynajmniej raz w przeciągu 4 lat. Jak widać, strategia doboru podmiotów do kontroli dotycząca tej grupy oparta jest na niezwykle prostym algorytmie oceny każdego podatnika. Scoring w tym przypadku oparty jest na wyliczeniu procentowego stosunku dochodu generowanego przez podatnika do całego dochodu z tytułu podatków. Agencja prowadzi scoring podatników zgodnie z omówionym algorytmem i ustala optymalny próg, na podstawie którego tworzona jest lista osób będących celem działań kontrolerów. Istnieją oczywiście pewne podobieństwa dla starej i nowej strategii kontroli podatkowej, jednak system przyznawania punktów poszczególnym podatnikom różni się zasadniczo dla każdego z rozwiązań. Dotychczasowa strategia jest jednowymiarowa i opiera się jedynie na ocenie wysokości odprowadzanego podatku. Nowy system doboru podmiotów do kontroli (Audit Select) bierze pod uwagę bardzo dużą liczbę cech (zob. rysunek 3.). Wyselekcjonowana w ten sposób grupa podatników jest węższa w porównaniu do dotychczasowego rozwiązania, a dodatkowo skuteczność kontroli w tej grupie jest znacznie wyższa (występująca liczba i wartość nadużyć podatkowych jest wyższa niż w przypadku starego rozwiązania).

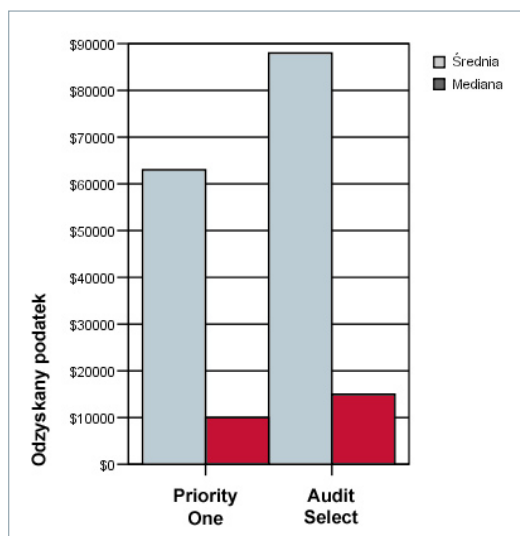
### Prior Productive (druga grupa podatników)

Do tej grupy trafiają wszyscy podatnicy, którzy we wcześniejszych zeznaniach popełnili błędy na kwotę większą niż \$10 000 (i oczywiście zostali na tym przyłapani). Podobnie jak w przypadku pierwszej grupy tak i w tej, kryterium kwalifikacji podatnika opiera się na jednym warunku. W przypadku systemu Audit Select kryterium to jest znacznie bardziej złożone i składa się z wielu warunków. Przeciętna korekta podatku w pierwszej grupie podatników (Priority One) wynosi \$76 000 (mediana \$9 600). Jest to wynik znacznie wyższy od \$12 000 (mediana \$1 300) w grupie pozostałych klientów (non-Priority One) oraz \$18 000 (mediana \$1 600) dla podatników ogółem. Trzeba zaznaczyć, że liczba kontroli podatkowych w grupie Priority One stanowi jedynie 9% wszystkich kontroli przeprowadzanych w stanie Teksas.

### Audit Select a Priority One

Kiedy porównujemy różne metody selekcji podatników do kontroli okazuje się, że ważne jest zwracanie uwagi nie tylko na przeciętny poziom korekt dla każdej metody, lecz także na liczbę kontroli. Jest to szczególnie istotne kiedy przeprowadzamy selekcję podatników w oparciu o modele scoringowe, takie jak użyte w Audit Select. Zastosowanie systemu oceny punktowej (scorigu) umożliwia stworzenie rankingu potencjalnych celów kontroli podatkowej, przyjęcie odpowiedniego przedziału punktowego spowoduje wybór określonej liczby podatników, którzy powinni zostać sprawdzeni. Chcąc porównać starą i nową strategię kontroli podatkowej trzeba za pomocą nowej metody (Audit Select) wybrać i sprawdzić grupę podatników, której liczebność jest zbliżona do grup wybieranych starą metodą (np. Priority One). W tym przypadku porównujemy wynik kontroli w grupie

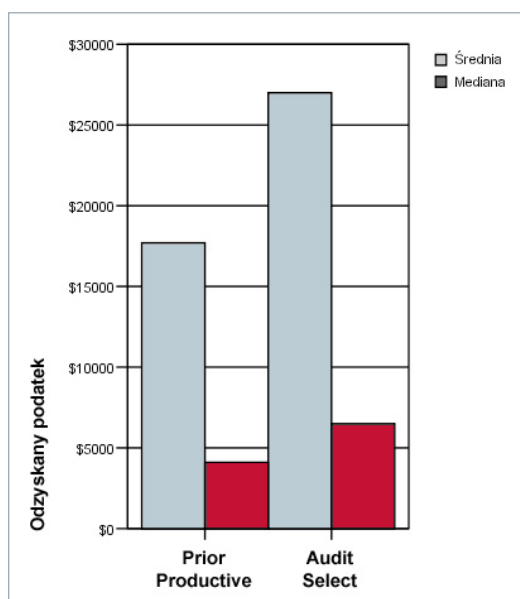
Priority One z wynikami kontroli w grupie wybranej na podstawie system selekcji Audit Select (za pomocą której wybrano podobną liczbę podatników). Przeciętna korekta podatku w grupie wybranej na podstawie nowej strategii wyniosła \$88 000 (mediana \$16 000) – zob. rysunek 5.). Wynik tej samej liczby kontroli jest o 16% wyższy niż uzyskany na podstawie starej metody selekcji. Oba systemy selekcji mogą zostać zintegrowane, wówczas nowa metoda Audit Select może służyć poprawie wyników metody Priority One poprzez dodatkową segmentację podatników i eliminację przypadków, dla których prawdopodobieństwo nadużycia jest niskie.



**Rysunek 5.**  
Porównanie metody Audit Select z Priority One

### Audit Select a Prior Productive

Tak jak w poprzednim przypadku, istnieje istotna różnica pomiędzy wynikami uzyskiwanymi za pomocą systemu Audit Select, a Prior Productive. Około 20% wszystkich kontroli związanych jest z kryterium selekcji podatników Prior Productive. 5% podatników w tej grupie mogłoby zostać zakwalifikowane do grupy Priority One. Przeciętne wyniki kontroli podatkowej w tej grupie (wyluczając wspomniane 5% tej grupy) kształtują się następująco: przeciętna korekta wyniosła \$17 700 (mediana \$4 100). Wybór tak samo licznej grupy podatników do kontroli metodą Audit Select (wyluczając przypadki, które potencjalnie znalazłyby się w grupie Priority One) skutkuje średnią korektą na poziomie \$27 000 (mediana \$6 500).



**Rysunek 6.**  
Porównanie metody Audit Select z Prior Productive

## Rozwijanie modelu

Rozwiązanie z zakresu drążenia danych jest szczególnie użyteczne, gdy jest integralną częścią systemu, bądź aplikacji. Proces drążenia danych nie jest procesem statycznym. Wraz ze zmianami zachodzącymi w danych lub procesach następuje jego ewolucja.

Od momentu wdrożenia systemu selekcji kontroli podatkowej Audit Select uległ on znacznym zmianom. Kilka czynników szczególnie przyczyniło się do rozwoju tego systemu. Najważniejszym jest stały wzrost jakości danych. Użyteczność i dokładność każdego modelu jest przede wszystkim zdeterminowana jakością, ilością i bogactwem dostępnych informacji. Drugim czynnikiem jest ciągle poszerzająca się wiedza na temat systemu podatkowego i jego związków z systemem gospodarczym. Ostatnim czynnikiem jest postęp w dziedzinie technik drążenia danych. Pojawiają się coraz bardziej wyrafinowane algorytmy modeli analitycznych i transformacji danych, które umożliwiają znaczące polepszenie wyników końcowych.

## Inne zastosowania drążenia danych w administracji podatkowej

Audit Select jest jednym z przykładów zastosowania metod drążenia danych w administracji podatkowej. Stosowane techniki analityczne powodują wzrost efektywności kontroli podatkowych i wskaźnika ściągalności podatków. Drążenie danych umożliwia również wykorzystanie i integrację różnorodnych źródeł danych, dostępnych we współczesnych hurtowniach. Poniżej znajduje się kilka przykładów zastosowania metod drążenia danych przez instytucje podatkowe.

### Selekcja obserwacji odstających

W tym podejściu drążenie danych zastosowano do budowania profili typowych zachowań podatników i wyboru tych, którzy do stworzonych profili nie pasują. Modelowanie rozpoczyna się od znalezienia charakterystycznych i jednorodnych segmentów podatników, a następnie stworzenie ogólnego wzorca podatnika dla każdego zidentyfikowanego segmentu. Na podstawie wzorców następuje selekcja przypadków odstających, które nie pasują do znalezionych reguł. Metoda ta jest podobna do stosowanych na co dzień w kontroli podatkowej, jednak w przypadku drążenia możliwe jest jednoczesne przetwarzanie ogromnej ilości danych na podstawie rozległych charakterystyk każdego podatnika.

### Profilowanie podatników zgodnie z typami działalności gospodarczej

Niektóre agencje kontrolujące podatników używają technik drążenia danych do analizy związków pomiędzy typem prowadzonej działalności gospodarczej, a rodzajem zeznania, które podatnik ma obowiązek wypełnić. Podobnie jak w przypadku budowania profili zachowań podatników, tak i w tym, możliwe jest wychwycenie charakterystycznych reguł rządzących związkami pomiędzy branżą lub typem działalności przedsiębiorstwa a rodzajem płaconych podatków. Uzyskane informacje mogą zostać wielokrotnie zastosowane na kompletnej bazie podatników, w celu wychwycenia przypadków ewentualnych nadużyć (prób niewypełnienia odpowiednich zeznań podatkowych).

### Model optymalnej selekcji

W dowolnej strategii kontroli, w której istotna jest identyfikacja docelowej grupy podatników zawsze musi dojść do kompromisu pomiędzy liczbą wyselekcjonowanych przypadków a ilością faktycznie przeprowadzanych kontroli. Techniki drążenia danych umożliwiają zastosowanie informacji na temat przeciętnego kosztu kontroli do wyznaczenia optymalnego zakresu działań. Ograniczenie liczby podatników i wybór szczególnie podejrzanych przypadków pozwala zminimalizować koszty, a tym samym maksymalizować efektywność przeprowadzanych kontroli.

## Wnioski

Drążenie danych ma szereg zastosowań w administracji podatkowej, a w szczególności w identyfikacji nadużyć podatkowych. W przypadku strategii selekcji zastosowanej w stanie Teksas, modele predykcyjne umożliwiły lepszą identyfikację osób i instytucji nie płacących należnych podatków. Pozwoliło to na pełniejsze wykorzystanie dostępnych zasobów ludzkich i materialnych oraz wzrost efektywności prowadzonych kontroli. W tym przypadku drążenie danych pomogło instytucji ulepszyć tradycyjne strategie kontroli, w celu osiągnięcia lepszych wyników.

## white PAPER

SPSS dostarcza wiedzę i narzędzia, które pozwalają na efektywną realizację projektów badawczych. Dostarczamy rozwiązania z zakresu zarządzania relacjami z klientem (CRM) i business intelligence, które umożliwiają użytkownikom systemów SPSS bardziej dochodową współpracę z ich klientami. Narzędzia SPSS pozwalają scalać i analizować dane marketingowe, dane o klientach i dane operacyjne w obrębie najważniejszych branż na całym świecie – między innymi w telekomunikacji, ochronie zdrowia, bankowości, finansach, ubezpieczeniach, produkcji, handlu, badaniach rynku, administracji, edukacji i sektorze publicznym. Poza centralą w Chicago (USA) SPSS posiada blisko 170 biur na całym świecie.

SPSS Polska zapewnia pełną informację o produktach SPSS, prowadzi kursy i szkolenia z zakresu analizy danych oraz obsługi i zastosowań programów SPSS. Użytkownikom zapewnia serwis i pomoc techniczną. Więcej informacji znajdą Państwo na stronach SPSS Polska, dostępnych pod adresem [www.spss.pl](http://www.spss.pl).

---

### **SPSS Polska**

ul. Raclawicka 58  
30-017 Kraków  
tel./faks 012.636.96.80  
tel./faks 012.636.07.91  
tel./faks 012.636.45.35  
e-mail: [info@spss.pl](mailto:info@spss.pl)  
[www.spss.pl](http://www.spss.pl)  
[www.analizadanych.pl](http://www.analizadanych.pl)  
[www.webmining.pl](http://www.webmining.pl)