IBM Analytics
Biała księga

Fundamentalna metodologia do analizy danych

IBM

W dziedzinie analizy danych rozwiązywanie problemów i udzielanie odpowiedzi na pytania przez analizę danych jest standardową praktyką. Często analitycy danych konstruują model do przewidywania wyników lub odkrycia bazowych wzorców w celu uzyskania prawidłowych wniosków. Organizacje mogą następnie wykorzystać te informacje do podjęcia działań, które doskonale poprawią przyszłe wyniki.

Istnieje wiele szybko rozwijających się technologii do analizy danych i budowania modeli. W zadziwiająco krótkim czasie przeszły ewolucję z komputerów stacjonarnych do masowo równoległych magazynów z ogromnymi zbiorami danych i analitycznymi funkcjami wewnątrz baz danych w relacyjnych bazach danych i platformie Apache Hadoop. Analiza tekstowa na niestrukturalnych lub półstrukturalnych danych staje się coraz istotniejsza jako sposób na wprowadzanie nastrojów i innych użytecznych informacji uzyskanych z tekstu do modeli predykcyjnych, co często prowadzi do znacznej poprawy jakości i dokładności modelu.

Pojawiające się podejścia analityczne dążą do automatyzacji wielu etapów budowy i zastosowania modelu, dzięki czemu technologie uczenia maszynowego stają się bardziej dostępne dla osób, które nie mają szerokich umiejętności analitycznych. Ponadto, w przeciwieństwie do podejścia „odgórnego", które polega na wstępnym definiowaniu problemu biznesowego, a następnie analizowaniu danych w celu znalezienia rozwiązania, niektórzy analitycy danych mogą wykorzystywać podejście „oddolne". Za pomocą tego ostatniego podejścia analityk danych uzyskuje wgląd w wielkie zbiory danych w celu znalezienia celu biznesowego, który może być zaproponowany na podstawie tych danych, a następnie rozwiązuje ten problem. W związku z tym, że większość problemów jest rozwiązywana w sposób odgórny, metodologia w niniejszym opracowaniu odzwierciedla ten pogląd.

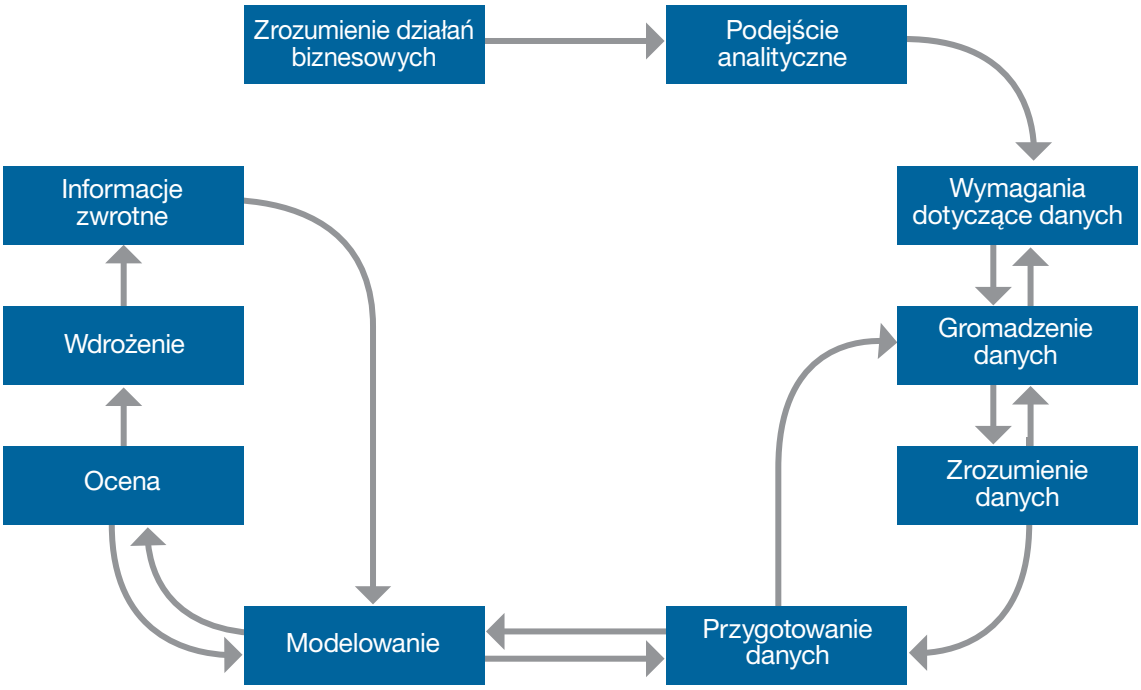
10-stopniowa metodologia analizy danych obejmująca technologie i podejścia

Ponieważ możliwości w zakresie analizy danych stają się coraz bardziej dostępne i powszechne, analitycy danych potrzebują fundamentalnej metodologii zdolnej do zapewnienia przewodniej strategii, niezależnie od zaangażowanych technologii, ilości danych lub podejść (patrz rys. 1). Metodologia ta wykazuje pewne podobieństwa do szeroko stosowanych metodologii[1–5] wyszukiwania danych, ale wprowadza kilka nowych rozwiązań w analizie danych, takich jak korzystanie z wielkich zbiorów danych, włączenie analizy tekstowej do modelowania predykcyjnego i automatyzacja niektórych procesów.

Powyższa metodologia składa się z 10 etapów, które tworzą proces iteracyjny w celu wykorzystania danych do uzyskania wniosków. Każdy etap odgrywa istotną rolę w kontekście ogólnej metodologii.

Czym jest metodologia?

Metodologia jest ogólną strategią kierującą procesami i działaniami w ramach danej branży. Metodologia nie zależy od konkretnych technologii lub narzędzi, nie jest to też zestaw technik i przepisów. Metodologia zapewnia analitykom danych ramy postępowania z dowolnymi metodami, procesami i heurystykami, które zostaną wykorzystane do uzyskania odpowiedzi lub wyników.



Rys. 1. Fundamentalna metodologia dla analizy danych

Etap 1. Zrozumienie działań biznesowych

Każdy projekt rozpoczyna się od zrozumienia działań biznesowych. Sponsorzy biznesowi, którzy potrzebują rozwiązania analitycznego, odgrywają najistotniejszą rolę na tym etapie przez zdefiniowanie problemu, celów projektu i wymagań dotyczących rozwiązań z perspektywy działalności biznesowej. Ten pierwszy etap stanowi podstawę dla pomyślnego rozwiązania problemu biznesowego. Aby zagwarantować sukces projektu, sponsorzy powinni być zaangażowani przez cały okres realizacji projektu w celu przekazania specjalistycznej wiedzy branżowej, dokonania przeglądu wyników pośrednich i zapewnienia, że prace pozostają na dobrej drodze do otrzymania zamierzonego rozwiązania.

Etap 2. Podejście analityczne

Po jasnym określeniu problemu biznesowego analityk danych może zdefiniować analityczne podejście do rozwiązania problemu. Ten etap wymaga wyrażenia problemu w kontekście technik statystycznych i uczenia maszynowego, tak aby organizacja mogła określić najwłaściwsze techniki dla osiągnięcia pożądanego rezultatu. Jeśli na przykład celem jest przewidywanie odpowiedzi takich jak „tak” lub „nie”, podejście analityczne można zdefiniować jako budowanie, testowanie i wdrożenie modelu klasyfikacyjnego.

Etap 3. Wymagania dotyczące danych

Wybrane podejście analityczne określa wymagania dotyczące danych. W szczególności stosowane metody analityczne wymagają określonej treści, formatów i reprezentacji danych w oparciu o wiedzę branżową.

Etap 4. Gromadzenie danych

Na początkowym etapie zbierania danych analitycy danych identyfikują i gromadzą dostępne źródła danych — strukturalne, niestrukturalne i półstrukturalne — istotne dla dziedziny problemu. Zazwyczaj muszą zdecydować, czy należy dokonać dodatkowych inwestycji w celu uzyskania mniej dostępnych elementów danych. Wówczas najlepszym wyjściem jest odroczenie decyzji inwestycyjnej do momentu uzyskania większych ilości informacji na temat danych i modelu. Jeśli istnieją luki w procesie gromadzenia danych, analityk danych może być zmuszony do odpowiedniej zmiany wymogów dotyczących danych i zebrania nowych danych i/lub większej ich ilości.

Podczas gdy próbkowanie danych oraz dzielenie ich na podzbiory są nadal ważnymi procesami, dzisiejsze wysokowydajne platformy i analityczne funkcje wewnątrz baz danych pozwalają analitykom danych na stosowanie znacznie większych zbiorów danych, zawierających dużą część lub nawet wszystkie dostępne dane. Dzięki zastosowaniu większej ilości danych modele predykcyjne mogą być w stanie lepiej opisywać rzadkie zdarzenia, takie jak częstość występowania choroby lub awaria systemu.

Etap 5. Zrozumienie danych

Po pierwotnym zebraniu danych analitycy danych zazwyczaj wykorzystują statystykę opisową i techniki wizualizacyjne, aby zrozumieć treść danych, ocenić jakość danych i sformułować podstawowe wnioski dotyczące danych. Dodatkowe gromadzenie danych może być konieczne w celu wypełnienia luk.

Etap 6. Przygotowanie danych

Etap ten obejmuje wszystkie działania mające na celu utworzenie zestawu danych, który będzie wykorzystywany w kolejnym etapie modelowania. Działania dotyczące przygotowania danych obejmują czyszczenie danych (w tym identyfikację wartości brakujących lub nieprawidłowych, eliminację duplikatów, prawidłowe formatowanie), łączenie danych z wielu źródeł (plików, tabel, platform) i przekształcanie danych w bardziej przydatne zmienne.

W procesie zwanym *inżynierią funkcjonalną* analitycy danych mogą tworzyć dodatkowe zmienne objaśniające, nazywane również *prognostykami* lub *funkcjami*, dzięki połączeniu wiedzy branżowej i istniejących zmiennych strukturalnych. Gdy dostępne są dane tekstowe, takie jak dzienniki biura obsługi klienta lub notatki lekarzy w niestrukturalnej lub półstrukturalnej formie, analiza tekstowa umożliwia uzyskanie nowych zmiennych strukturalnych w celu wzbogacenia zbioru prognostyków oraz poprawy dokładności modelu.

Przygotowanie danych jest zazwyczaj najbardziej czasochłonnym etapem w projekcie analizy danych. W wielu branżach niektóre etapy przygotowania danych są wspólne dla różnych problemów. Wstępna automatyzacja niektórych etapów przygotowania danych może przyspieszyć ten proces przez zminimalizowanie doraźnego czasu przygotowania. Dzięki dzisiejszym wysokowydajnym, masowo równoległym systemom wysokiej wydajności i funkcjom analitycznym zaimplementowanym w miejscach przechowywania danych analitycy danych mogą łatwiej i szybciej przygotować dane przy użyciu bardzo dużych zbiorów danych.

Etap 7. Modelowanie

Począwszy od pierwszej wersji przygotowanego zbioru danych etap modelowania skupia się na rozwijaniu modeli predykcyjnych lub opisowych zgodnie z wcześniej zdefiniowanym podejściem analitycznym. Wykorzystując modele predykcyjne, analitycy danych używają zestawu *szkoleniowego* (danych historycznych, dla których znany jest wynik finansowy) w celu zbudowania modelu.

Proces modelowania jest zazwyczaj wysoce iteracyjny, gdyż organizacje uzyskują pośrednie wnioski, co prowadzi do udoskonaleń w przygotowaniu danych i specyfikacji modelu. Dla danej techniki analitycy danych mogą testować wiele algorytmów z ich odpowiednimi parametrami, aby znaleźć najlepszy model dla dostępnych zmiennych.

Etap 8. Ocena

Podczas opracowywania modelu i przed jego wdrożeniem analityk danych ocenia dany model, aby wyznaczyć jego jakość i upewnić się, że prawidłowo i całkowicie rozwiązuje on problem biznesowy. Ocena modelu wymaga obliczenia różnych działań diagnostycznych i innych wartości wyjściowych, takich jak tabele i wykresy, umożliwiających analitykowi danych interpretację jakości modelu oraz jego skuteczności w rozwiązywaniu problemu. W przypadku modelu predykcyjnego analitycy danych używają zestawu testowego, który jest niezależny od zestawu szkoleniowego, ale ma taki sam rozkład prawdopodobieństwa i posiada znany wynik. Zestaw testowy jest używany do oceny modelu, w związku z czym może być w razie potrzeby udoskonalany. Niekiedy ostateczny model jest również stosowany do zestawu walidacji do celu dokonania ostatecznej oceny.

Dodatkowo analitycy danych mogą przypisać testy istotności statystycznej do modelu jako kolejne potwierdzenie jego jakości. To dodatkowe potwierdzenie może odgrywać zasadniczą rolę w uzasadnieniu wdrożenia modelu lub podjęciu działań o wysokim ryzyku, takich jak kosztowny dodatkowy protokół medyczny lub krytyczny system sterowania lotem samolotu.

Etap 9. Wdrożenie

Po opracowaniu satysfakcjonującego modelu i jego zatwierdzeniu przez sponsorów biznesowych jest on wdrażany do środowiska produkcyjnego lub podobnego środowiska testowego. Zazwyczaj jest wdrażany w ograniczonym zakresie do czasu, gdy jego działanie zostanie w pełni przetestowane. Wdrożenie może być tak proste jak wygenerowanie raportu z zaleceniami lub tak skomplikowane jak osadzenie modelu w złożonym procesie przepływu pracy i procesu oceny zarządzanych przez aplikację użytkownika. Wdrożenie modelu do operacyjnego procesu biznesowego zazwyczaj angażuje dodatkowe grupy, umiejętności i technologie w ramach przedsiębiorstwa. Na przykład grupa sprzedażowa może wdrożyć model odpowiedzi na tendencje za pośrednictwem procesu zarządzania kampanią stworzonego przez zespół rozwoju i administrowany przez grupę marketingową.

Etap 10. Informacje zwrotne

Przez gromadzenie wyników z wdrożonego modelu organizacja otrzymuje informacje zwrotne dotyczące wydajności modelu i jego wpływu na środowisko, w którym został wdrożony. Informacje zwrotne mogą na przykład przybrać formę wskaźników odpowiedzi na kampanię promocyjną skierowaną do grupy klientów wskazanych przez model jako respondenci o wysokim potencjale. Analiza tych informacji zwrotnych pozwala analitykom danych udoskonalenie modelu w celu poprawy jego dokładności i użyteczności. Mogą zautomatyzować niektóre lub wszystkie etapy gromadzenia informacji zwrotnych i oceniania modelu, udoskonalania i powtórnego jego wdrażania w celu przyspieszenia procesu odświeżania modelu dla uzyskania lepszych wyników.

Zapewnianie trwałej wartości dla organizacji

Przepływ metodologii ilustruje iteracyjny charakter procesu rozwiązywania problemów. W związku z tym, że analitycy danych gromadzą więcej informacji na temat danych i modelowania, często wracają do poprzedniego etapu w celu dokonania zmian. Modele nie są tworzone jednorazowo, wdrażane i pozostawiane w stanie, w jakim się znajdują; zamiast tego, dzięki sprzężeniu zwrotnemu, udoskonalaniu i powtórnemu wdrażaniu, modele są stale udoskonalane i dostosowywane do zmieniających się warunków. W ten sposób zarówno model, jak i składające się na jego otrzymanie działania mogą stanowić trwałą wartość dla organizacji tak długo, jak to rozwiązanie jest potrzebne.

Więcej informacji

Nowy kurs dotyczący fundamentalnej metodologii analizy danych jest dostępny w Big Data University. Bezpłatny kurs online jest dostępny pod adresem: http://bigdatauniversity.com/bdu-wp/bdu-course/data-science-methodology

Robocze przykłady wdrożenia tej metodologii w przypadkach rzeczywistego użycia można znaleźć pod adresem:

- http://ibm.co/1SUhxFm
- http://ibm.co/1IazTvG

Podziękowania

Dziękuję Michaelowi Haidemu, dr. Michaelowi Wurstowi, Brandonowi MacKenziemu i Gregory'emu Roddowi za ich cenne uwagi oraz Jo A. Ramosowi za jego rolę w rozwoju tej metodologii podczas wielu lat naszej współpracy.

O autorze

Dr John B. Rollins jest analitykiem danych w organizacji IBM® Analytics. Ma doświadczenie w inżynierii, wyszukiwaniu danych i ekonometrii w wielu branżach. Jest właścicielem siedmiu patentów oraz autorem bestsellerowego podręcznika inżynierii i wielu artykułów specjalistycznych. Ma stopień doktora z dziedziny inżynierii naftowej i ekonomii Teksańskiego Uniwersytetu Rolniczego i Mechanicznego.

IBM Polska
ul. 1 sierpnia
02-134 Warszawa

[1] Brachman R. i Anand T., „The process of knowledge discovery in databases" w: Fayyad U. et al., red., Advances in knowledge discovery and data mining, AAAI Press, 1996 (s. 37–57)

[2] SAS Institute, http://en.wikipedia.org/wiki/SEMMA, www.sas.com/en_us/software/analytics/enterprise-miner.html, www.sas.com/en_gb/software/small-midsize-business/desktop-data-mining.html

[3] Wikipedia, „Cross Industry Standard Process for Data Mining", http://en.wikipedia.org/wiki/Cross_Industry_Standard_Process_for_Data_Mining, http://the-modeling-agency.com/crisp-dm.pdf

[4] Ballard C., Rollins J., Ramos J., Perkins A., Hale R., Dorneich A., Milner E. i Chodagam J., Dynamic Warehousing: Data Mining Made Easy, IBM Redbook SG24-7418-00 (wrzesień 2007), s. 9–26

[5] Gregory Piatetsky, CRISP-DM, still the top methodology for analytics, data mining, or data science projects, 28 października 2014, www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html

IMW14828-PLPL-00