Data science – nowa nauka w obszarze analizy danych

Data science jest nową, powstałą niespełna kilka lat temu, dziedziną nauki. Obecnie uważa się ją za podstawową i wiodącą dyscyplinę w obszarze budowy modeli prognostycznych.

W kolejnych artykułach tego cyklu omawiać będę poszczególne metody prognozowania i klasyfikacji stosowane przy określaniu przyszłych cen zbóż i przewidywaniu tendencji na rynkach surowcowych.

Czym jest *data science*?

Data science jest dziedziną nauki łączącą wiele dyscyplin, takich jak matematyka, statystyka, ekonometria i informatyka. Najważniejszym bodźcem do jej powstania był dynamiczny rozwój systemów przetwarzania danych oraz towarzyszący jej ogromny wzrost ilości dostępnych danych. Dzięki rozwojowi informatyki opracowane kiedyś teoretyczne metody prognozowania i modelowania „ożyły" i stały się możliwe do praktycznego wykorzystania.

Kiedyś, bez komputerów stosowanie złożonych modeli sieci neuronowych lub drzew decyzyjnych wiązało się z długotrwałymi, żmudnymi obliczeniami. Pracochłonność i czas obliczeń oraz brak łatwych w użyciu danych stanowiły skuteczną barierę w rozpowszechnianiu się nowatorskich metod analitycznych w obszarze biznesu.

Termin *data science* pojawił się w różnych kontekstach dopiero w ostatnich latach. Przełomowym momentem w powstaniu tej dziedziny był, opublikowany w 2012 roku w czasopiśmie „Harvard Business Review", artykuł amerykańskiego badacza danych Dhanurjay Patila. W artykule „Data Scientist: The Sexiest Job of the 21st Century" stwierdził, że specjalista ds. danych jest „nową rasą", a „brak naukowców danych staje się poważnym ograniczeniem w niektórych sektorach". Dalej opisał również zakres nowej nauki i jej znaczenie w gospodarce przyszłości. Po roku 2017 pojęcie *data science* stało się powszechnie znane na świecie. Istnieje zatem w przestrzeni publicznej zaledwie od trzech lat.

Dlaczego pojawił się *data science*?

Data science jest bardzo silnie związana z technologią zbierania i przetwarzania danych. Jej powstanie związane jest nie tylko z dynamicznym rozwojem informatyki, lecz przede wszystkim z dramatycznym wzrostem ilości danych. Kiedyś procesy gospodarcze generowały tylko skromne, podstawowe dane. Wraz ze wzrostem złożoności procesów przemysłowych i transformacji gospodarki nastąpił lawinowy wzrost dostępnych informacji. Każda maszyna, każdy pojazd i system sprzedażowy produkują dzisiaj ogromne ilości danych w postaci tekstowych szeregów czasowych. Część danych ma fundamentalne znaczenie przy modelowaniu, optymalizacji procesów i tym samym osiąganiu przewagi konkurencyjnej. Pojawiła się więc biznesowa potrzeba efektywnego wykorzystywania danych.

W nowej sytuacji zastosowanie tradycyjnych narzędzi analitycznych okazało się niepraktyczne. Używane powszechnie przez analityków arkusze kalkulacyjne okazały się bezsilne w obliczu ogromnych ilości danych i złożoności procesów. Z kolei narzędzia używane dotąd w obszarze analiz statystycznych okazały się mało elastyczne i niezdolne do szybkiej adaptacji nowych rozwiązań analitycznych.

Środowiska programistyczne w *data science*

Analizy *data science* mogą być prowadzone w arkuszach kalkulacyjnych. Niestety powszechnie używane arkusze typu Excel są w stanie efektywnie przetworzyć tekstowe bazy danych do wielkości 0,2 GB. Niestety, większość modeli prognostycznych operuje na znacznie większych bazach danych tekstowych z zakresu od 0,1 do nawet 30 GB.

Duży zakres danych wymusił więc odejście od tradycyjnego arkusza kalkulacyjnego w kierunku środowiska programistycznego.

Podstawowym formatem baz danych *data science* są pliki tekstowe, w tym najpopularniejsze standardy: csv i txt. Powód jest bardzo prosty: większość urządzeń i systemów w otaczającej gospodarce generuje dane w postaci tekstowych tabel płaskich. *Data science* jest nauką badawczą, nie specjalizuje się w systemach bazodanowych i rzadko używa relacyjnych baz danych.

Głównymi językami programowania w d*ata science* są R i Python.

Język R został opracowany przez Johna Chambersa w Bell Laboratories w roku 2000. W większości napisany jest w języku C oraz Fortrain. Powszechnie uważa się, że z powodu Fortraina jest on wolniejszy od konkurencyjnego Pythona, który w całości został napisany w języku C. Python został napisany we wczesnych latach 90. przez amerykańskiego naukowca Guido van Rossum. Nazwa języka pochodzi od serialu komediowego emitowanego w latach 70. przez BBC „Monty Python's Flying Circus" („Latający cyrk Monty Pythona").

Język R stosowany jest częściej w ośrodkach akademickich, natomiast Python częściej można spotkać w centrach analitycznych. Obecnie zauważalna jest tendencja do odchodzenie od języka R na rzecz języka Python.

Języki programowania używane są w specjalnych edytorach, spośród których najbardziej popularne są edytory: Jupyter notebook oraz Pycharm.

Powszechnie uważa się, że najlepszym systemem operacyjnym do pracy z językami R i Python jest system Ubuntu. Oczywiście języki te mogą pracować na innych systemach operacyjnych takich jak MS Windowsk czy Mac OS.

Przy tworzeniu zaawansowanych modeli matematycznych mocno obciążających zasoby obliczeniowe stosuje się m.in środowiska wirtualne Apache: Hadoop, Spark, Kafka, Cassandra, Hive i inne. Systemy te najczęściej stosowane są w dużych ośrodkach badawczych i centrach analitycznych.

Zarówno języki R i Python podobnie jak ich edytory: Jupyter i Pycharm oraz system operacyjny Ubuntu są dostępne za darmo w sieci Internet.

Biblioteki w *data science*

Mimo że *data science* jest bardzo mocno osadzona w językach programowania, praca analityczna polega nie na programowaniu w tych językach, lecz na wykorzystaniu specjalistycznych bibliotek.

Biblioteki są swoistymi subjęzykami wyspecjalizowanymi do określonych zadań. Każda biblioteka ma swój własny zestaw komend i specyficzną składnię. Najczęściej składnia bibliotek jest podobna do składni ich języków bazowych Python lub R. Powszechnie uważa się, że Python ma znacznie bogatsze biblioteki od języka R.

Najważniejszymi bibliotekami stosowanymi w Python w obszarze sieci neuronowych są: TensorFlow oraz ostatnio zyskujący na popularności Pytorch. Przy analizach graficznych stosuje się biblioteki: Seaborn, Matplotlib oraz Plotly. Modele prognostyczne obsługiwane są przez biblioteki: Sklearn, Statmodels oraz Numpy. Niewątpliwie ważniejszą biblioteką jest też Pandas. Nazwa pochodzi od skrótu „panel data". Biblioteka ta służy do pobierania i przygotowywania danych do użycia w innych bibliotekach.

Biblioteki podobnie jak inne narzędzia pracy *data science* są bezpłatne.

Jakie zmiany przynosi *data science*?

Nowa dziedzina nauki jaką jest *data science* wprowadza szereg rewolucyjnych zmian, które na zawsze zmieniły oblicze prowadzenia badań i analiz. Z punktu widzenia analityków istnieją trzy najważniejsze zmiany:

1. Wzrost znaczenia języków programowania i ich bibliotek

Jest to bardzo duża zmiana, ponieważ do tej pory tego typu prace wykonywana były na drogich programach prognostycznych typu: Minitab, Statistica lub SAS.

W odróżnieniu od wspomnianych systemów analitycznych biblioteki podlegają częstym modyfikacjom, w których wdrażane są najnowsze osiągnięcia nauki. Biblioteki mają zdolność wzajemnej współpracy, przez co są bardziej elastyczne w działaniu w odróżnieniu od używanych dotychczas programów.

2. Wzrost dostępności wiedzy oraz jej aktualność i szybkie jej przekazywanie

Obecnie wiedza, w postaci praktycznych porad, samouczków i gotowych skryptów, dostępna jest praktyczne od ręki w Internecie. Co tydzień pojawiają się nowe rozwiązania matematyczne i programistyczne, które często fundamentalnie zmieniają sposób prowadzenia badań i tworzenia modeli. Pojawiło się zjawisko przedwczesnej dezaktualizacji książek, jeszcze przed ich wydrukowaniem.

3. Praca w chmurze analitycznej

Wzrost intensywności procesów przetwarzania danych doprowadził do powstania nowego rodzaju usług polegających na pracy zdalnej z wykorzystaniem wydzierżawionej mocy obliczeniowej komputerów zewnętrznych. Jest to zupełnie nowe zjawisko posiadające ogromny potencjał rozwojowy.

XXI wiek jest nazywany wiekiem informacji. *Data science* jest obszarem działalności, który idealnie wkomponowuje się w obecne czasy.

Wojciech Moszczyński