Jakie dane są potrzebne, aby zastosować sztuczną inteligencję?

Zastosowanie modeli ekonometrycznych, wszelkiego rodzaju rozwiązań i algorytmów związanych z optymalizacją i przewidywaniem przyszłości staje się coraz bardziej popularne w środowiskach biznesowych. Przedsiębiorcy przewidując nadchodzącą recesję i chcą się jak najlepiej przygotować. Szukają więc rozwiązań mogących zwiększyć efektywność prowadzonej działalności, zaoszczędzić koszty, zwiększyć rentowność. Przedsiębiorcy rzadko orientują się w metodach optymalizacji i obszarach zastosowania sztucznej inteligencji. Jest to całkiem zrozumiałe, ponieważ metody te są dość skomplikowane zarówno w obszarze ich działania, jak i zagadnień związanych ze stworzeniem odpowiedniego środowiska dla zastosowania tych metod. Przedsiębiorcy więc szukają kontaktu z firmami, które specjalizują się implementacji tego typu rozwiązań. Pierwszym i najważniejszym krokiem przy rozpoczęciu takiej współpracy jest zdefiniowanie potrzeb i oczekiwań przedsiębiorcy oraz omówienie źródeł danych i jakości danych posiadanych przez przedsiębiorstwo.

Czym są dane dla data science?

Dane dla analityków danych są rzeczą fundamentalną. Ich jakość, ilość oraz wiarygodność są podstawą wszelkich analiz i bezpośrednio przekładają się na jakości rozwiązań, jakie analitycy ci będą zdolni dostarczyć. Dla data science dane są jak paliwo dla samochodu. Dane są bezkompromisowo najważniejsze. Dlatego zawsze na samym początku firmy specjalizujące się w dostarczaniu rozwiązań data science analizują jakość danych. Potrzeby klientów to jedno, a możliwości ich zaspokojenia to drugie. Często zdarza się, że klienci potrzebują rozwiązań, które są niemożliwe do zbudowania, ponieważ nie istnieją odpowiednie dane historyczne.

Dane nietypowe

Jak pamiętamy data science to nauka, która operuje w najróżniejszych obszarach życia i działalności gospodarczej. Modele data science potrafią identyfikować twarze, odciski palców, potrafią na podstawie dźwięku rozróżniać poszczególne osoby, mogą wyłapywać zmiany w sygnaturze cieplnej wielkich urządzeń lub analizować ruchy klientów w trakcie zakupów w supermarkecie. Wymienione tutaj zdolności nie są jednak często wykorzystywane w środowisku przedsiębiorców. Najbardziej typowymi potrzebami biznesu są działania w obszarze liczbowym zawartym w systemach typu ERP. Tak więc gdybyśmy wyłączyli z obszaru data science takie źródła danych jak zdjęcia, filmy czy nagrane dźwięki to pozostanie obszar danych typowo ilościowych i jakościowych. W dzisiejszych czasach dane zbierane są praktycznie przez każde bardziej złożone urządzenie. Każdy samochód osobowy zbiera jednocześnie kilkadziesiąt rozmaitych danych, każdy smartfon zbiera dane na temat lokalizacji, wykonanych połączeń, zużycia baterii czy poboru danych z chmury. Wszystkie te dane trafiają do historii urządzeń. Dane historyczne są najważniejsze źródłem danych dla rozwiązań data science.

Cztery obszary danych

W typowych przedsiębiorstwach można spotkać cztery rodzaje danych. Pierwszy rodzaj to rejestr sprzedaży.

Rejestr sprzedaży

Jest to zazwyczaj bardzo wysokiej jakości baza, która zawiera informacje o czasie dokonania transakcji, wartości dokonanej transakcji, jej przedmiocie oraz często wskazuje osoby, która uczestniczy w transakcji, wskazuje osoby kupujące. Rejestr sprzedaży nie zawiera błędów, braków i jest bardzo wiarygodny. Związane jest to z tym, że rejestr ten jest tworzony przez kasę fiskalną, która jest bardzo mocno obwarowana prawnie i jest potencjalnym przedmiotem kontroli urzędów skarbowych. Taka baza ma postać szeregu czasowego, każda transakcja jest po kolei rejestrowana, a głównym elementem spajającym jest czas zawierania transakcji. Baza może być wykorzystana zarówno do budowania modeli predykcyjnych, jak i modeli zachowań klientów, wykrywania anomalii, tendencji i trendów.

Baza przebiegu procesu

To również doskonała baza danych dla zastosowań data science. Niestety nie jest ona już tak niezawodna jak rejestr sprzedaży. Baza składa się z szeregów czasowych dokonywanych pomiarów w procesach przemysłowych. Na przykład jeżeli mamy silos zbożowy to system może zbierać informacje na temat wilgotności wewnątrz magazynu, temperatury, ciśnienia lub zapylenia. Pomiar wszystkich tych parametrów może odbywać się jednocześnie i odkładać się w bazie danych co godzinę. Powstaje więc spójny zestaw szeregów czasowych zapisujących historię warunków panujących w silosie zbożowym w ujęciu czasowym. Problem polega na tym, że czasami niektóre czujniki się psują, inne pokazują przekłamane informacje. Generalnie dane tego typu są w ograniczonym zakresie wiarygodne i niestety nie zawsze kompletne. Awaria jednego z czujników eliminuje odczyt z tego czujnika na wiele dni lub tygodni, co pogarsza możliwości tworzenia modeli prognostycznych wychwytujących anomalie lub przewidujących pewne zjawiska, których chcemy uniknąć.

Baza magazynowa

Kolejnym typowym źródłem danych jest baza opisująca stany magazynowe. Baza ta również ma charakter szeregów

czasowych i może mieć podobną dynamikę jak rejestr sprzedaży. Dane zawarte w takich bazach mogą okazać się również częściowo niewiarygodne. Możemy zaobserwować to, gdy inwentaryzacja ujawnia istotne różnice inwentaryzacyjne. Bazy tego typu wykorzystuje się do modeli, które mają za zadanie optymalizować strukturę magazynową, zwiększać rotację magazynową oraz przeciwdziałać nadmiernemu zamrażaniu kapitału pracującego. Dane magazynowe mają też istotne znaczenie przy tworzeniu optymalnych systemów zarządzania łańcuchem dostaw.

Baza charakterystyk

Charakterystyki to kolejny, ostatni, najczęściej spotykany typ danych. Są to np. spisy charakterystyk towarów, gdzie możemy znaleźć informacje o wielkości towaru i wymaganiach składowania towaru. Mogą to być też spisy cech poszczególnych klientów, informacje na temat terminów napraw urządzeń lub wymaganych terminów przeglądów technicznych. Bazy charakterystyk są bardzo istotne przy tworzeniu modeli zachowań. Dzięki indywidualnym cechom przedmiotów, urządzeń lub klientów możliwe jest, w połączeniu z bazami przebiegu procesu lub ewidencją sprzedaży, stworzenie modeli, które będą wychwytywały zależność przyczynowo-skutkową pomiędzy cechami a wynikami. Modele są matematyczną interpretacją rzeczywistości, są w pewnym sensie rodzajem uogólnienia procesu i mają niezwykłe zdolności kojarzenia zależności zjawisk. Dzięki informacji o wieku lub płci klientów model jest w stanie przyporządkować pewne zachowania lub skłonności w kontekście tych cech. Nie ma znaczenia jak wiele będzie cech, modele z nieprawdopodobną sprawnością będą znajdowały zależności pomiędzy cechami a wartościami wynikowymi. Wszystko to ma ogromne znaczenie w procesie budowania rozwiązań data science.

Wymagania związane z danymi potrzebnymi do tworzenia rozwiązań data science

Aby dane można było wykorzystać muszą one spełniać określone własności podstawowe. Poniżej przedstawiłem sześć najważniejszych cech, jakie trzeba ocenić przy poborze danych.

1. Dane muszą być wiarygodne

Niewiarygodnych danych nie można wykorzystać w analizach data science.

Klasycznym przykładem niewiarygodnych danych są dane zbierane na podstawie wypełnionych przez klientów lub pracowników deklaracji, formularzy i innych form, które nie są weryfikowalne. W takich formularzach klient lub pracownik prowadzący badania może wpisać każde dane. Informacje te najczęściej są niemożliwe do weryfikacji. Dla kontrastu, przykładem danych stuprocentowo wiarygodnych są te rejestrowane przez maszyny lub rejestry sprzedaży tworzony przez kasy fiskalne.

2. Dane muszą być kompletne

Niekompletne dane można wykorzystać w analizach data science. Nawet jeżeli uda się skompletować informacje za pomocą algorytmów data science w niekompletnej bazie danych, informacje te nie będą do końca wiarygodne. Innym sposobem radzenia sobie z niekompletnymi danymi jest kasowanie rekordów tam, gdzie dane są niekompletne. To również jest rozwiązaniem niedoskonałym, ponieważ braki danych mogą występować w określonych sytuacjach, a skasowanie rekordów z brakami, mając na uwadze, że braki te związane były z określonymi sytuacjami, prowadzi do osłabienia modeli lub nawet uniemożliwi tworzenie modeli matematycznych.

3. Dane mogą składać się z wielu odrębnych baz danych

Aby wykorzystać dane w analizach data science należy zespolić różne bazy danych w jedną zintegrowaną bazę. Jeżeli istnieje konieczność zespalania danych z różnych baz danych konieczne jest istnienie w poszczególnych bazach kluczy identyfikacyjnych. Dzięki nim możliwe jest wzajemne łączenie wielu baz. Tak więc, jeżeli w jednej bazie występują dane klienta i kluczem tej bazy będzie ID klienta to możliwe jest połączenie klienta z rejestrem sprzedaży właśnie według klucza ID, który występuje zarówno w rejestrze sprzedaży, jak i w bazie klientów.

4. Dane powinny mieć format liczbowy lub format kategoryczny

Możliwe są również inne formaty, takie jak np. baza komentarzy tekstowych. Dane takie wymagają zastosowania algorytmów rozpoznawania treści. Algorytmy te przetwarzają tekst na kategorie, czyli na format kategoryczny. Najważniejszymi formatami danych ilościowych są: formaty liczbowy i kategoryczny. Oba mogą występować jednocześnie w jednej bazie. Format kategoryczny tworzony jest przez kategorie. Kategoriami może być np. płeć: 1-kobieta, 2-mężczyzna, dzień miesiąca od 1 do 31, dzień tygodnia, kolor oczu lub kategorie 1- słońce i 0 – brak słońca. Jeżeli dane zjawisko można zapisać w postaci liczb całkowitych i liczba unikalnych wystąpień tych liczb w zbiorze nie przekracza 100 sztuk możemy mówić o tym, że są to dane kategoryczne. Mogą być one zapisywane cyfrowo lub literowo. Danymi kategorycznymi są na przykład kategorie rozmów zapisywane w formularzach biura obsługi klienta.

5. Bazy powinny mieć postać plików płaskich

Akceptowalnymi formatami dla data science są pliki w formacie arkuszy kalkulacyjnych Excela lub LibraOffice oraz formaty tekstowe: txt i csv. Istnieje możliwość pobierania danych na potrzeby analizy data science z relacyjnych baz danych. Konieczna jest wtedy ze strony analityków danych implementacja bibliotek SQL. Dostęp do takich informacji jest możliwy poprzez przekazanie przez właściciela bazy kluczy dostępowych lub udostępnienie aplikacji dostępowej API do pobierania danych.

6. Dane w postaci szeregu czasowego

Najbardziej pożądaną formą danych wykorzystywanych w analizach rynkowych są szeregi czasowe. To dane, które są uporządkowane według czasu, gdzie w jednej z kolumn bazy znajduje się data i godzina zdarzenia, np. sprzedaży lub dostawy. Szeregi czasowe umożliwiają wykrywanie trendów sprzedaży, martwych okresów i innych zjawisk o charakterze cyklicznym ściśle powiązanym z kalendarzem. Niektóre

dane, takie jak np. rejestr klientów, rejestr sklepów czy charakterystyki produktów, nie mają charakteru szeregów czasowych. Typowym szeregiem czasowym jest rejestr sprzedaży tworzony przez kasę fiskalną.

W procesach przemysłowych często korzysta się z szeregów czasowych utworzonych z odczytu narzędzi pomiarowych. Przykładem takiego szeregu jest zapis mierzonej co kilka minut temperatury.

Podsumowanie

Jakość danych jest determinantą zakresu wykorzystania nowoczesnych rozwiązań związanych z rozwojem transformacji cyfrowej. Wysiłki związane z pomiarami i składowaniem informacji mogą okazać się daremne, gdy dane nie będą spełniały wypisanych w tym artykule standardów. Zazwyczaj przedsiębiorstwa rozwijają swoje zasoby danych wraz z własnym bardziej lub mniej dynamicznym rozwojem. Budowanie takiego systemu danych przypomina spontaniczny rozrost rośliny, krzaku lub drzewa. Spontanicznie powstają kolejne bazy, kolejne nowe elementy w bazach pojawiają się w odpowiedzi na potrzeby rozwojowe przedsiębiorstwa. Dobrą praktyką jest audyt jakości danych oraz jakości całej architektury składowania i poboru danych. Jeżeli system danych jest nieuporządkowany, niespójny, mówimy o istnieniu długu technologicznego. Aby zastosować nowoczesne metody ilościowe musimy najpierw uporządkować dane, spowodować zwiększenie ich spójności i wiarygodności. Działania takie nie mają zazwyczaj wpływu na poprawę efektywności, oszczędzania kosztów czy zwiększenia rentowności produkcji. Dlatego działanie te nazywane są spłacaniem długu technologicznego.

Wojciech Moszczyński

Wojciech Moszczyński – Ekspert z zakresu optymalizacji matematycznej oraz modelowania predykcyjnego. Od lat zaangażowany w popularyzację metod ekonometrycznych w środowisku biznesowym. Specjalizuje się w optymalizacji procesów sprzedażowych, produkcyjnych oraz logistycznych. Przez 15 lat pracował jako ekspert finansowy ze specjalizacją w obszarze kontrolingu i rachunkowości zarządczej. Od 10 lat pracuje jako analityk danych (*data scientist*). Absolwent Katedry Ekonometrii i Statystyki Uniwersytetu Mikołaja Kopernika w Toruniu. Obecnie zatrudniony jako Senior Data Scientist w polskiej firmie Unity Group.