



Modele matematyczne w data science

Tworzenia optymalizacji i wdrażania rozwiązań z zakresu sztucznej inteligencji opiera się przede wszystkim na tworzeniu modeli matematycznych. Modele matematyczne to uproszczenie rzeczywistości bazujące na danych historycznych. Dane te wzajemnie się przenikają, są wewnętrznie powiązane, współzależne od siebie. Modele wykorzystują te zależności, znajdują najczęściej występujące, powtarzające się zjawiska i tworzą swego rodzaju makiety rzeczywistości, rodzaj sprzężenia zwrotnego zachowań opartych na rachunku prawdopodobieństwa. Przedsiębiorca, który planuje wdrożenie rozwiązań z obszaru sztucznej inteligencji nie musi znać się na budowaniu modeli matematycznych. Warto jednak, aby z grubsza wiedział czym są modele, jakie są ich najważniejsze cechy, a przede wszystkim jakie powinny istnieć warunki, aby takie modele można było budować.

Dane potrzebne do budowy modeli matematycznych

Każde przedsiębiorstwo ma bazy danych. Mogą to być rejestry sprzedaży, ewidencje środków obrotowych w magazynach, rejestry klientów lub informacje związane z przepływem surowców. Są to bazy wewnętrzne. Do tworzenia modeli wykorzystuje się również bazy zewnętrzne, takie jak: historyczne dane pogodowe, informacje o nastrojach makroekonomicznych czy historyczne przebiegi poziomu aktywności klientów na rynku. Wyróżniamy pięć podstawowych rodzajów baz danych występujących niemal w każdym przedsiębiorstwie, które są wykorzystywane do budowy modeli matematycznych. Są to bazy:

- CRM,
- ewidencyjne,
- finansowe,
- transakcji (paragonowe),
- przebiegu procesu.

Baza CRM jest to rejestr przebiegu współpracy z klientami. To zapisy historyczne terminów wysłanych ofert, spotkań, notatki z negocjacji związanych z konkretnymi klientami.

Bazy ewidencyjne mają charakter stacjonarny. Zawierające stałe informacje na temat określonych obiektów, takich jak kubatura, adres czy odległość. Bazy te mogą zawierać również informacje stałe na temat klientów takie jak wiek, płeć, adres prowadzenia działalności czy wykształcenie. Mogą zawierać też adresy sklepów detalicznych, zdolności produkcyjne określonych maszyn lub krańcowe możliwości przesyłowe określonych instalacji.

Bazy finansowe to zbiory danych zawierające informacje dotyczące przychodów, kosztów i wyników finansowego uzyskiwanych w określonych miejscach powstawania kosztów (tzw. MPK). Bazy takie mają firmy, które używają rozwiązań z zakresu rachunkowości zarządczej.

Bazy transakcji (zwane również bazami paragonów) są to na przykład rejestry sprzedaży na kasach fiskalnych w sklepach detalicznych, rejestry wysyłek i przyjęć towarów

w magazynach. Bazy takie mają format szeregów czasowych, ponieważ ich głównym elementem jest czas dokonywania transakcji. Znakomicie nadają się do wytyczania trendów, wyszukiwania zależności i powiązań z czynnikami zewnętrznymi.

Bazy przebiegu procesów mają podobny format szeregów czasowych jak bazy transakcji. To rejestry oparte na czasie obserwacji, są więc szeregiem czasowymi. Odzwierciedlają przebieg zarejestrowanego procesu według określonych zjawisk fizycznych lub chemicznych zachodzących w instalacjach produkcyjnych lub transportowych.

Każda firma prowadząca złożoną działalność gospodarczą ma przynajmniej jedną z pięciu wymienionych tutaj baz danych.

Tworzenie modelu matematycznego

Jak już wspomniałem model matematyczny to uproszczenie rzeczywistości. Pierwszym krokiem przy tworzeniu modelu jest określenie jego celu. Najczęściej modele tworzy się w celu przewidywania przyszłości, rzadziej wykorzystuje się je do interpretowania bieżących zjawisk. Załóżmy, że chodzi nam o przewidywanie jaka będzie cena tony mąki na rynku za dwa miesiące. Należy zebrać informacje, jaki był poziom cen w kolejnych tygodniach w ciągu ostatnich trzech, czterech lat. Zestawiając takie informacje zauważymy, że ceny mąki zmieniały się w powtarzających się cyklach, powstawały pewne trendy, w niektórych miesiącach cena mąki była niższa, w innych raptownie wzrastała tworząc maksima okresowe. Możemy zaobserwować, że na przykład, zawsze najwyższa cena określonego surowca występowała w marcu i kwietniu. To pierwsza obserwacja, proste uogólnienie rzeczywistości, które można wykorzystać biznesowo. Dokonałmy go bez pomocy modelu, po prostu patrząc na wykres. Wiemy o tym, że na cenę mąki wpływ miał nie tylko miesiąc, ale również pogoda, ceny walut, ogólny stan gospodarki, poziom konsumpcji oraz lokalny wolumen importu zboża. Wiele zjawisk miało bardziej lub mniej subtelny wpływ na ceny surowca, niektóre zjawiska wpływały na ceny z dużym opóźnieniem. Zjawiska takie jak wielkość eksportu, koszty nawozów i ceny paliwa są ze sobą ściśle skorelowane tworząc złożone kombinacje zależności. Powiązania te wyrażane są przez rachunek prawdopodobieństwa, relacje najczęściej są przesunięte w czasie, dlatego nie można ich wykryć bez wykorzystania metod statystycznych.

Przeciętnemu człowiekowi bez pomocy technologii przetwarzania danych trudno jest powiązać wszystkie fakty. Do interpretacji złożonych danych służy model matematyczny. Człowiek musi

w roku, to, że stosunkowo wysokim prawdopodobieństwem, możemy określić jaką będzie cena surowca w maju tego roku. Ale jak łatwo zauważyć wtedy nie możemy bazować wyłącznie na interpolacji! I bez nas musimy uruchomić jeden z prostych modeli autogresyji średniej ruchomej ARMA lub ARIMA. To dość kłopotliwy sposób, dlatego można skorzystać z innej metody.

Nauka z nadzorowaniem i walidacja modeli!

Opisując najważniejsze techniki modelowania data science trzeba wspomnieć czym jest uczenie nadzorowane. To prosta metoda, która polega na tym, że do trenowania modelu wykorzystujemy zazwyczaj 70% danych, którymi dysponujemy. Pozostałe 30% danych służą do tego, aby przetestować jakość stworzonych prognoz. W przypadku naszego modelu prognozującego ceny maki założymy, że dysponujemy 36 miesiącami danych historycznych. Do trenowania modelu wykorzystujemy 30 miesięcy. Pozostałe sześć wykorzystujemy do testowania modelu. Testowanie prognoz polega na sprawdzaniu jak precyzyjnie model prognozuje ceny po podstawieniu zmiannych opisujących z wspomnianych, ostatnich sześciu miesięcy. Model tworzy prognozy, a my znając jakie były ceny w tych ostatnich sześciu miesiącach możemy porównać o ile prognozujemy różniły się od prawdziwych cen. Działania to nazywamy walidacją modelu. Dzięki walidacji modelu możemy dowiedzieć się, czy modele przewidyuje trafnie, czy jednak konieczne są dalsze prace nad poprawą jego skuteczności.

Podsumowanie

Modelowanie jest najważniejszą techniką tworzenia rozwiązań z obszaru sztucznej inteligencji. Modele możliwe są do tworzenia przede wszystkim w oparciu o dane historyczne. Modele matematyczne są uproszczeniem rzeczywistości, istotne jest, aby uproszczenie to nie wpływało na jakiś dokonanych prognoz lub interpretacji zjawisk. Przedsiębiorca, który zlecił budowanie modeli matematycznych musi mieć świadomość, że na końcową trafność prognoz modeli wpływ ma przede wszystkim jakość danych, ich wiarygodność oraz kompletność. Dlatego ważna dla przyszłości przedsiębiorstwa jest troska o jakość zbierania i przechowywania danych.

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

wskazać, które dane, zwane zmiennymi niezależnymi, będą miały charakter szeregów opisyjących, a które dane będą zmiennymi wynikowymi. Jeżeli budujemy model przewidujący cenę maki w przyszłości zmienną wynikową będzie historia historyczna cena maki w poszczególnych miesiącach.

Zagregowane dane w postaci szeregów czasowych zawierające zmienne opisowe są zestawiane w algorytmie modelu matematycznego ze zmiennymi wynikowymi. Model sam znajduje zależności i wyciąga konkretne wnioski. W efekcie, na podstawie historycznych zmiennych opisyjących, powsta- je uproszczenie rzeczywistości, którego wynikiem jest prognoza cen maki w przyszłości.

Rodzaje modeli matematycznych

Podstawowym podziałem modeli matematycznych wykorzystywanych w technikach data science są modele klasyfikacji i modele regresji. Modele klasyfikacji dają wynik w tzw. formacie dyskretnym. To informacja, czy będzie padło (wynik oznaczony jako: 1), czy nie będzie padło (wynik oznaczony jako: 0). Model klasyfikacji może na przykład przewidywać, który rodzaj klientów będzie negatywnie reagował na zainicjowane zmiany w sprzedaży. Tak więc model może na przykład przewidzieć, że najbardziej spontaniczne reakcje spośród klientów będą występowały w grupie wiekowej 30-45 lat (wynik oznaczony jako grupa: 0). Jednocześnie model wskazuje, że najbardziej odporne na grupą klientów będzie grupa klientów w wieku 18-25 lat (wynik oznaczony jako grupa: 1).

Zupełnie inny charakter przekazywanych informacji mają modele regresji. Wskazują one precyzyjnie jaka będzie przyszła prognozowana wielkość. Tak więc model może powiedzieć, że za 10 dni cena surfowca może kosztować 175 zł za tonę lub, że temperatura jutro o godzinie 12:00 będzie wynosiła 14,5 stopnia Celsjusza.

W jaki sposób modele przewidują przyszłość?

jak już wspominałem modele bazują na uogólnieniach utworzonych na podstawie obserwacji danych historycznych. Dane te mają określone wielkości, ich zestawienie tworzy pewne uogólnienie prowadzące do uzyskania wyniku. W jaki sposób model może stworzyć prognozę przyszłości skoro nie wiemy jaki będzie poziom wartości zmienionych opisujących w przyszłości? Istnieją dwa rozwiązania, które mogą działać jednocześnie. Pierwszy sposób to wykazywać stanie trendów. Zmienne opisujące mogą wykazywać pewne trendy historyczne, dzięki czemu możliwa jest interpolacja ich wielkości w przyszłości. Na podstawie zachowania i trendów historycznych możliwa jest interpolacja wielkości zmienionych opisujących. Jeżeli więc w okresie czterech lat zawsze w maju cena paliwa, będąca zmienną opisującą, była najwyższą i cena ta stanowiła na przykład 22% przeciętnej ceny