

# Zastosowanie modelu Incremental Isolation Forest do poprawy jakości produkcji spożywczej



Wojciech Moszczyński

Isolation Forest to model machine learning, który może być stosowany do wykrywania anomalii na taśmie produkcyjnej żywności, np. pakującej mięso. Taka instalacja ma szereg czujników, które nadsyłają informacje w sposób ciągły. Odbiór ich na komputerze osobistym jest stosunkowo prosty. Można skierować strumień danych bezpośrednio z czujników do chmury analitycznej np. Azure lub AWS, a tam można uruchomić gotowy model, który będzie wysyłał wszelkie ostrzeżenia związane wykrytymi anomaliami w produkcji.

**T**akie rozwiązanie jest łatwe, proste (dla przeciętnego analityka danych) i często darmowe. Załóżmy, że mamy dwa sensory pomiarowe na taśmie pakującej mięso:

- czujniki wagi, które mierzą czy każdy kawałek ma odpowiednią wagę,
- czujniki temperatury, które mierzą temperaturę mięsa, co może świadczyć o jego świeżości.

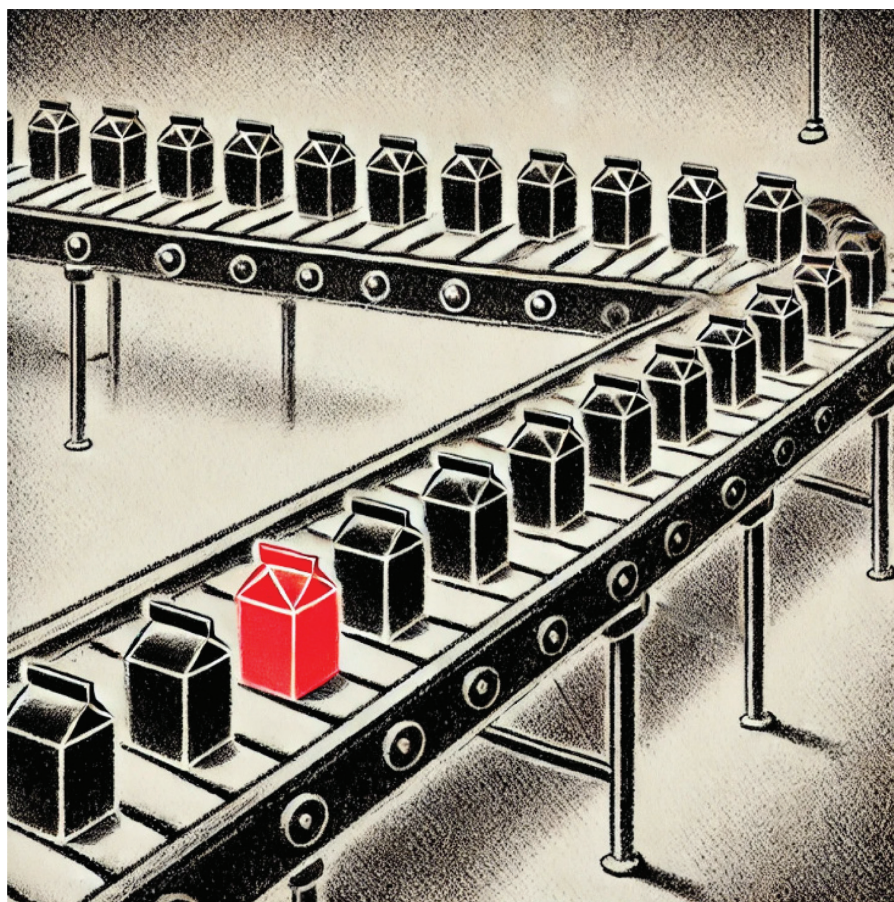
Dzięki temu dla każdego produktu zbierany jest zestaw danych: waga i temperatura.

Model Isolation Forest uczy się na podstawie wielu produktów, które przeszły przez taśmę. W ten sposób wie, jak wygląda „typowy” produkt. Nagle może pojawić się produkt nietypowy, np. zbyt ciężki kawałek mięsa albo zbyt ciepły, co może świadczyć, że mięso zaczyna się psuć. Wtedy taki produkt zostanie rozpoznany przez Isolation Forest jako anomalia.

Dlaczego tak się dzieje? Bo produkt, który odstaje od reszty (ma inną wagę lub temperaturę) będzie szybko izolowany przez drzewa decyzyjne modelu. Modelowi wystarczy kilka podziałów, by zauważyć, że ten produkt „nie pasuje” do standardowych produktów, które normalnie pojawiają się na taśmie.

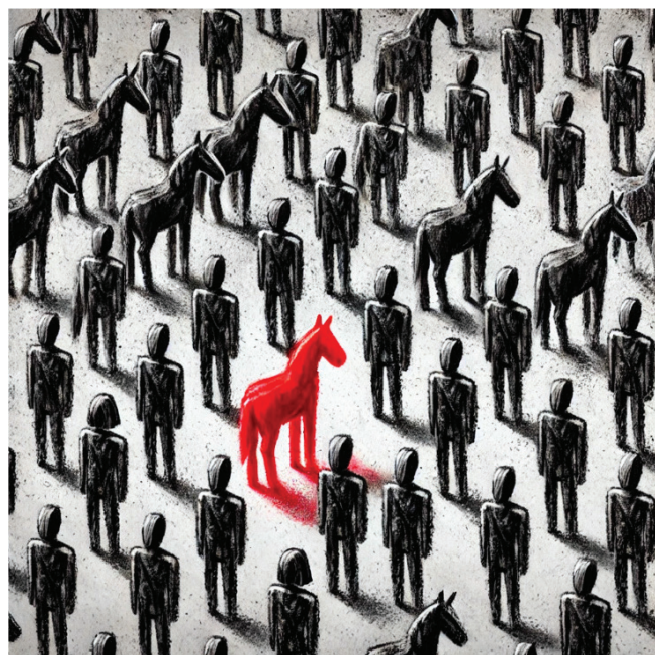
## Jak działa model Isolation Forest

Model Isolation Forest jest używany do wykrywania anomalii w grupie elementów. Identyfikuje więc coś, co nie pasuje do reszty. Model buduje drzewa losowo, wybór cech jest przypadkowy (np. płeć klientów banku, rasa, wiek, dochód klientów). A więc kształt drzew nie jest zaplanowany.



Z bazy klientów drzewo może losowo dzielić np. najpierw na płeć (kobiety i mężczyźni), później miejsce zamieszkania, wiek, itd. Jeśli nagle pojawi się coś zupełnie nietypowego (koń, kot), to taki punkt danych będzie bardzo szybko odseparowany (po jednym lub dwóch podziałach) – utworzy się pojedyncza, krótka gałąź drzewa. I właśnie to wskazuje, że ten punkt jest anomalia.

Warto podkreślić, że kluczowe jest właśnie szybkie odizolowanie punktów anomalnych (np. w moim przykładzie zwierząt od ludzi). Identyfikacja anomalii odbywa się przez identyfikację



niewielkiej liczby gałęzi. Niewielka liczba gałęzi wynika z tego, że zwierzęta w moim przykładzie nie mają wspólnych cech z resztą grupy klientów banku.

Każde drzewo powstaje tak, że losowo dzieli dane na coraz mniejsze grupy. Takie losowe dzielenie powtarza się wielokrotnie. Normalne obserwacje są podobne do siebie, więc żeby je od siebie oddzielić, trzeba wykonać wiele podziałów – powstaje dużo „gałęzi”. Anomalie są wtedy, gdy po kilku podziałach nie ma już możliwości dalszych podziałów. Gałąź znajduje się całkiem sama i krótka.

Czynnikiem, który decyduje o tym czy dana obserwacja jest anomalią, jest liczba gałęzi (podziałów) potrzebnych do odizolowania tej obserwacji od pozostałych punktów.

Isolation Forest jest dobry do analizy danych streamingowych z kilku powodów.

- Po pierwsze model ten jest bardzo szybki, buduje proste drzewa losowo i nie wymaga złożonych obliczeń. Dzięki temu przetwarza dane w czasie rzeczywistym, bez opóźnień. Nie wymaga przechowywania całych danych historycznych ani skomplikowanych wzorców.
- Idealnie nadaje się więc do szybkiego wykrywania pojedynczych nietypowych punktów pojawiających się na bieżąco.
- Ponadto model ten łatwo aktualizować. Ponieważ jest to model bez nadzoru elastycznie dostosowuje się do nowych danych, co jest niezbędne przy zmieniających się strumieniach danych.
- Nie wymaga dużej ilości pamięci, ponieważ drzewa nie korzystają z danych historycznych, korzysta tylko z historycznej struktury drzew.

Isolation Forest jest szybki, lekki obliczeniowo i dobrze nadaje się do szybkiego wykrywania anomalii „w locie” podczas przetwarzania danych streamingowych. Model ten w swojej podstawowej wersji analizuje dane statyczne (czyli zestawy danych już wcześniej zgromadzonych), ale można go też dostosować do analizy danych dynamicznych (strumieniowych).

## Nauka nadzorowana i nienadzorowana

W data science, czyli nauce o danych rozróżnia się analizy statyczną i dynamiczną. Analiza statyczna jest oparta o pewnym określonym zbiorze danych. Ten zbiór zawiera dane historyczne. Model uczy się na danych historycznych i kiedy już jest przetrenowany, można go zastosować do danych dynamicznych. Takie podejście nazywamy nauką z nadzorem, ponieważ nauka taka opiera się na zmiennych opisujących, zmiennych wpływających na wartość wyników.

Uczenie z nadzorem przypomina trochę sytuację, w której uczy się dziecko rozpoznawać zwierzęta pokazując mu zdjęcia i wyraźnie mówiąc: „to jest kot, a to jest pies”. Tutaj model dokładnie wie, jaki wynik powinien uzyskać dla danych przykładów, bo ma jasno określoną zmienną wynikową, czyli to, co próbujemy przewidzieć. Dzięki temu model może nauczyć się, jak cechy (np. wielkość zwierzęcia, długość sierści, dźwięk) wpływają na wynik (czy to kot czy pies).

Natomiast uczenie nienadzorowane wygląda inaczej. To sytuacja, gdy pokazuje się dziecku dużo zdjęć różnych zwierząt, ale w ogóle nie mówi się mu, jak się nazywają. Dziecko musi samo znaleźć podobieństwa lub różnice między nimi i na tej podstawie stworzyć jakieś grupy lub kategorie. W takim podejściu nie mamy zmiennej wynikowej, takiej jak lew, kot, koń – model sam próbuje zrozumieć strukturę danych i podzielić je na logiczne grupy albo wykryć różnice pomiędzy zwierzętami. Model ma sam wykryć grupy zwierząt, gdy jakieś stworzenie nie pasuje do żadnej z grup zwierząt, zostaje uznane jako anomalia.

Podsumowując, kluczowa różnica to fakt, że w nauce z nadzorem model dostaje jasną wskazówkę, jaki wynik powinien uzyskać, podczas gdy w uczeniu nienadzorowanym musi sam znaleźć wzorce i porządek w zbiorze danych, bo nie ma wcześniej określonych poprawnych odpowiedzi.



## Analiza statyczna lub dynamiczna (streamingowa)

W analizie statycznej (batchowej) mamy już gotowy, zamknięty zbiór danych (np. dane transakcji klientów z całego roku). Jednorazowo budowane są drzewa decyzyjne. W przypadku analizy dynamicznej (streamingowej) punkty napływają w czasie rzeczywistym. Nie ma stałego, dużego zbioru danych, tylko ciągle napływające nowe dane, np. transakcje w sklepie online czy dane z czujników instalacji produkcyjnych.

Isolation Forest potrzebuje danych historycznych do zbudowania losowego lasu drzew decyzyjnych. Następnie nowe napływające punkty sprawdza względem już stworzonych drzew, oceniając czy są anomaliami. Tym samym model nie musi przechowywać historycznych danych, wystarczy, że przechowuje strukturę drzew. Drzewa są uproszczeniem cech wcześniej analizowanych danych.

### Jak Isolation Forest działa dla procesów streamingowych?

Dla danych streamingowych, Isolation Forest jest stosowany w formie tzw. Incremental Isolation Forest. Na podstawie początkowych danych budowane jest pierwsze drzewo. Następnie, gdy napływają nowe dane w procesie strimingowym, model na bieżąco aktualizuje drzewa. Dzięki temu cały czas „uczy się” nowych wzorców i szybko reaguje na zmiany. Czyli jeżeli wrócimy do naszego pierwszego przykładu, gdy do zbioru klientów banku (tam gdzie pojawił się koń jako anomalia) dołączymy zbiór z kotami, dodamy dużo kotów, model potraktuje to jako grupę i zacznie ją identyfikować pod kontem koloru futerka, płci i wieku. Na bieżąco stworzy więc nowe drzewa. Tak więc w praktyce dla danych dynamicznych model Isolation Forest, potrafi dynamicznie adaptować się do strumienia danych.

### Podsumowanie

Podsumowując, modele Isolation Forest stanowią efektywne narzędzie do automatycznej detekcji anomalii, ponieważ mogą jednocześnie analizować setki różnych zmiennych, na podstawie których szybko i skutecznie identyfikują odstępstwa od normy. Zastosowanie takich modeli nie ogranicza się jedynie do produkcji taśmowej; mogą być one wykorzystywane również do analizy produktów przechowywanych w magazynach, surowców przyjmowanych na wagach wjazdowych w zakładach przetwórczych, a także do monitorowania procesów pomocniczych, takich jak wykorzystanie zasobów ludzkich czy systemu wsparcia produkcji.

Co istotne, po wdrożeniu do infrastruktury chmurowej, modele Isolation Forest działają całkowicie automatycznie i praktycznie bezobsługowo. Oznacza to, że raz wdrożone mogą



funkcjonować przez nieograniczony czas, a ich podatność na awarie czy błędy techniczne jest minimalna, co dodatkowo zwiększa niezawodność całego procesu kontroli jakości i monitorowania produkcji.

Implementacja praktyczna tego rozwiązania polega na podłączeniu strumienia danych z taśmy produkcyjnej bezpośrednio do chmury analitycznej poprzez interfejs API. Dane ze strumienia są następnie automatycznie przetwarzane przez moduł analityczny, oparty na modelu Incremental Isolation Forest, który na bieżąco klasyfikuje produkty jako „normalne” lub „anomalne”. Wyniki tej analizy są przysyłane w formie szeregów czasowych, które zawierają jasną i czytelną klasyfikację każdego analizowanego produktu. Tego typu powiadomienia mogą być przysyłane bezpośrednio na urządzenia mobilne, np. telefon komórkowy kierownika produkcji lub pracowników inspekcji nadzoru technicznego. Dzięki temu możliwa jest natychmiastowa reakcja na ewentualne problemy jakościowe lub technologiczne, co znacząco zwiększa efektywność kontroli jakości całego procesu produkcyjnego.

Wojciech Moszczyński

**Wojciech Moszczyński** – absolwent Katedry Ekonometrii i Statystyki Uniwersytetu Mikołaja Kopernika w Toruniu; specjalista z zakresu ekonometrii, finansów, data science oraz rachunkowości zarządczej. Specjalizuje się w optymalizacji procesów produkcyjnych i logistycznych. Prowadzi badania w obszarze rozwoju i zastosowania sztucznej inteligencji. Od lat zaangażowany w popularyzację machine learning oraz data science w środowiskach biznesowych.