4. UCZENIE MASZYNOWE (MACHINE LEARNING)

Autor: Dejan Mirčetić

Istnieje wiele pytań na temat tego, czym jest Uczenie Maszynowe (ang. Machine Learning – ML). Czy jest to rzeczywiście proces, w którym maszyny samodzielnie się uczą wykorzystując do tego otoczenie zewnętrzne, czy też jest to sformalizowany za pomocą algorytmów matematycznych proces, który pozwala komputerom „wymyślać" reguły funkcjonujące w świecie zewnętrznym? Jakie narzędzia wykorzystuje ML? Jak wygląda typowy przepływ danych w kanale ML? Czy jest on stosowalny w tradycyjnych branżach, a nie tylko w branżach związanych z IT i Internetem? Jakie miejsce ML zajmuje w kontekście biznesu? Jak systematycznie wykorzystywać je do rozwiązywania problemów biznesowych? Czy istnieje jakaś architektura dotycząca tego, jak stosować uczenie maszynowe w ŁD?

Na te i podobne pytania postaramy się udzielić odpowiedzi w kolejnym rozdziale, kończąc go rzeczywistym przykładem studium przypadku zastosowania algorytmów ML w łańcuchu dostaw żywności.

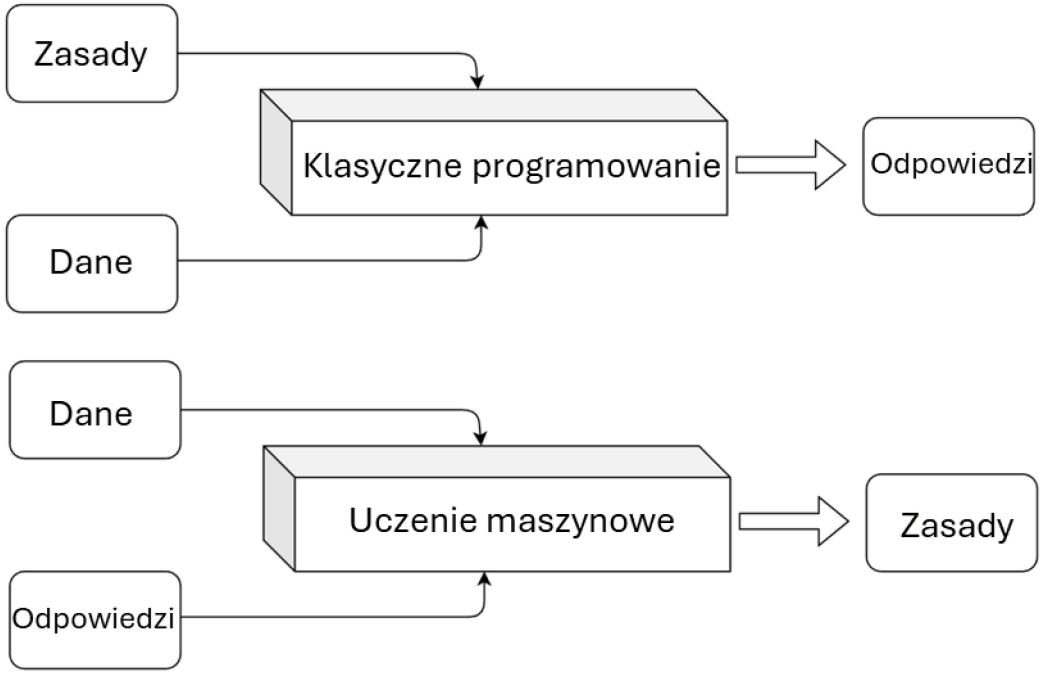
4.1. Czym jest uczenie maszynowe?

Uczenie maszynowe (ML) to dyscyplina skupiająca się na dwóch powiązanych ze sobą pytaniach: Jak można konstruować systemy komputerowe, które automatycznie ulepszają swoje możliwości wykorzystując do tego doświadczenie? oraz Jakie są podstawowe statystyczne prawa teorii informacji obliczeniowej, które rządzą wszystkimi systemami uczącymi się, w tym komputerami, ludźmi i organizacjami? Badanie uczenia maszynowego jest ważne zarówno dla znalezienia odpowiedzi na te podstawowe pytania naukowe i inżynieryjne, jak i dla wysoce praktycznego oprogramowania komputerowego, które zostało wyprodukowane i wdrożone w wielu aplikacjach (Jordan & Mitchell, 2015).

ML wynika z pytania: czy działanie komputera mogłoby wyjść poza „to, co wiemy, jak mu zlecić wykonanie jakiegoś zadania" i nauczyć się samodzielnie, jak wykonać określone czynności? Czy komputer mógłby nas zaskoczyć? Czy komputer mógłby automatycznie nauczyć się tych reguł, opierając się na danych, zamiast o ręcznie napisane

reguły przetwarzania danych stworzone przez programistów? To pytanie otwiera drzwi do nowego paradygmatu programowania (Chollet, 2021).

ML umożliwia fundamentalną zmianę paradygmatu programowania (Rysunek 4.1). W programowaniu klasycznym programista (człowiek) wprowadza reguły (program) oraz dane, które są analizowane i przetwarzane zgodnie z tymi regułami. W rezultacie odpowiedzi są dostarczane na końcu procesu. Z drugiej strony, w ML programista (człowiek) wprowadza dane z odpowiedziami oczekiwanymi od danych, a dopiero na ich podstawie powstają reguły.



Rysunek 4.1 Klasyczne programowanie kontra szkolenie systemów uczenia maszynowego

Źródło: Chollet (2021)

Klasyczne programowanie można rozumieć jako programowanie imperatywne, ponieważ programista wstępnie definiuje wszystkie reguły, a wykonywanie kodu odbywa się zgodnie z nimi, podczas gdy uczenie maszynowe można rozumieć jako programowanie deklaratywne, w którym wyrażamy cele wyższego poziomu lub opisujemy ważne ograniczenia i polegamy na algorytmach matematycznych, aby zdecydować, jak i/lub kiedy przełożyć to na działanie.

Obecnie ML jest podstawą niezliczonych istotnych dla wielu procesów aplikacji, w tym wyszukiwania w sieci, ochrony poczty e-mail przed spamem, rozpoznawania mowy, rekomendacji produktów i innych (Ng, Andrew 2017). Wielu programistów systemów AI zdaje sobie obecnie sprawę, że w przypadku wielu aplikacji o wiele łatwiej jest „szkolić" system, pokazując mu przykłady pożądanego zachowania wejścia-wyjścia, niż programować go

ręcznie, przewidując pożądaną odpowiedź dla wszystkich możliwych danych wejściowych. Efekt wykorzystania ML był również szeroko identyfikowany w całej branży informatycznej i w szeregu branż zajmujących się problemami intensywnie wykorzystującymi dane, takimi jak usługi konsumenckie, diagnostyka usterek w złożonych systemach i kontrola łańcuchów logistycznych (Jordan & Mitchell, 2015).

4.2. Podstawy i założenia teoretyczne ML

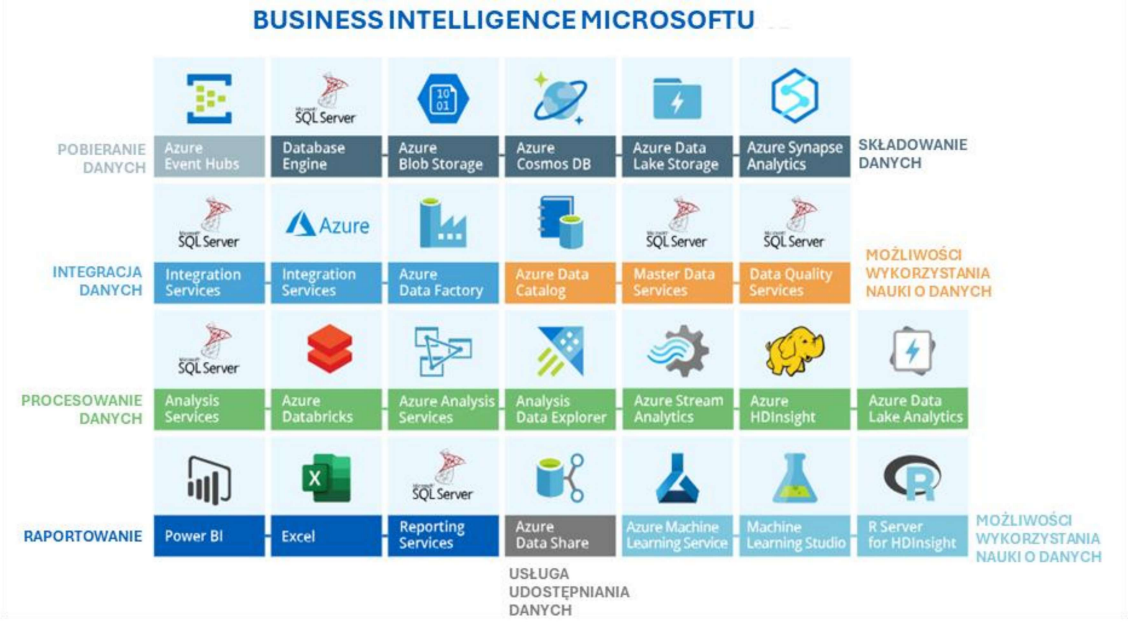
Pochodzenia ML należy szukać w matematyce, a dokładniej w statystyce. Stąd też ML wykorzystuje podstawy teoretyczne i algorytmy opracowane w **uczeniu statystycznym**, jednakże eksperci toczą dyskusję, czy ML jest prawdziwą dziedziną samą w sobie, czy też jest po prostu częścią statystyki. W praktyce algorytmy ML zwykle nie mają pewnego poziomu sztywności matematycznej i czasami łatwo przekraczają pewne ograniczenia matematyczne obecne w statystyce. Na przykład algorytmy ML nie zwracają zbytniej uwagi na przedziały ufności podczas optymalizacji współczynników w algorytmach parametrycznych, chociaż jest to jeden z najważniejszych tematów w statystyce. **Ogólnie rzecz biorąc, ML i statystyka w dużym stopniu się pokrywają**, a niektórzy z najbardziej znanych twórców i profesorów algorytmów ML twierdzą, że jest to po prostu część statystyki (Hastie i in., 2009). Niemniej jednak, będąc dziedziną samą w sobie lub częścią statystyki, ML składa się z kilku kroków w pozyskiwaniu wiedzy z danych. Nie istnieje jednak powszechny konsensus co do tych kroków, ale ogólnie rzecz biorąc, można je przedstawić jako transformację różnych źródeł danych w spostrzeżenia Business Intelligence.

W kontekście biznesowym modele ML są bezużyteczne, bez odpowiedniego wsparcia w zakresie wstępnego przetwarzania danych, eksploracji danych i stosowania spostrzeżeń w rzeczywistych procesach. Dlatego tworzenie algorytmów ML bez możliwości aktualizacji modelu i wykorzystania jego wyników do rzeczywistego procesu podejmowania decyzji nie przynosi żadnej wartości nowoczesnym przedsiębiorstwom. W związku z tym w dzisiejszej analityce biznesowej ilościowy proces ML jest zwykle częścią przepływu pracy Business Intelligence. Dokładniej rzecz biorąc, jest częścią ważnych podprocesów Business Intelligence (nauka o danych i analiza danych stanowią część Business Intelligence). Szczegóły dotyczące roli ML w tych procesach i samych rzeczywistych procesów generowania wartości dla biznesu za pośrednictwem ML zostaną przedstawione w kolejnym podrozdziale.

4.3. Business Intelligence i ML in ŁD

Business Intelligence, w kontekście ŁD, to proces wyciągania wniosków na temat obserwowanych procesów łańcucha dostaw, w oparciu o modelowanie danych generowanych w ramach tych procesów. BI opiera się głównie na statystyce, ale bierze również pod uwagę inne obszary matematyczne: badania operacyjne, algebra liniowa, logika rozmyta (w przypadku, gdy danych jest niewiele lub ich brakuje), optymalizacja numeryczna, metaheurystyka itp. Ponadto nowe technologie przełomowe stają się również ważnym aspektem analizy danych i dostarczania wniosków: **uczenie maszynowe, sztuczna inteligencja, cyfrowe bliźniaki, smartization, żywe laboratoria** itp.

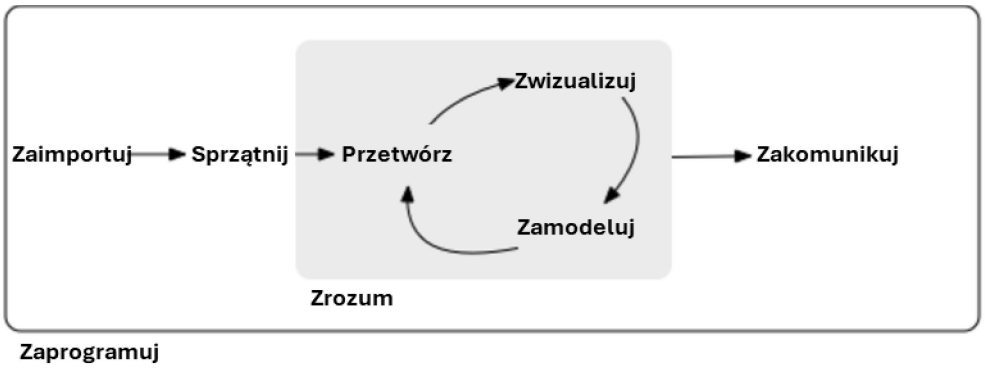
Nie ma ściśle zorganizowanych procedur dotyczących tego, jak powinny być zorganizowane procedury Business Intelligence i przepływy pracy ML, ale istnieją pewne przydatne wytyczne w praktyce i literaturze, które okazały się skuteczne podczas przeprowadzania analizy. Procedura przeprowadzania Business Intelligence jest również zróżnicowana w zależności od źródła oprogramowania, które jest używane do analizy. Na przykład Microsoft oferuje kilka narzędzi za pośrednictwem swojego pakietu Microsoft Business Intelligence, które wykonują różne zadania**:** **pobieranie danych, przechowywanie danych, integrację danych, zarządzanie danymi, przetwarzanie danych, raportowanie, udostępnianie danych oraz naukę o danych** (Rysunek 4.2).



Rysunek 4.2 Architektura Microsoft Business Intelligence

Źródło: ScienceSoft (b.d.)

W danej architekturze procedury ML są stosowane tylko na poziomie nauki o danych za pośrednictwem kilku narzędzi: usług Azure ML, ML studio i R Server dla HDInsight. Ogólna procedura wykonywania analizy danych w kontekście ML w R Server, jest przedstawiona na Rysunku 4.3.

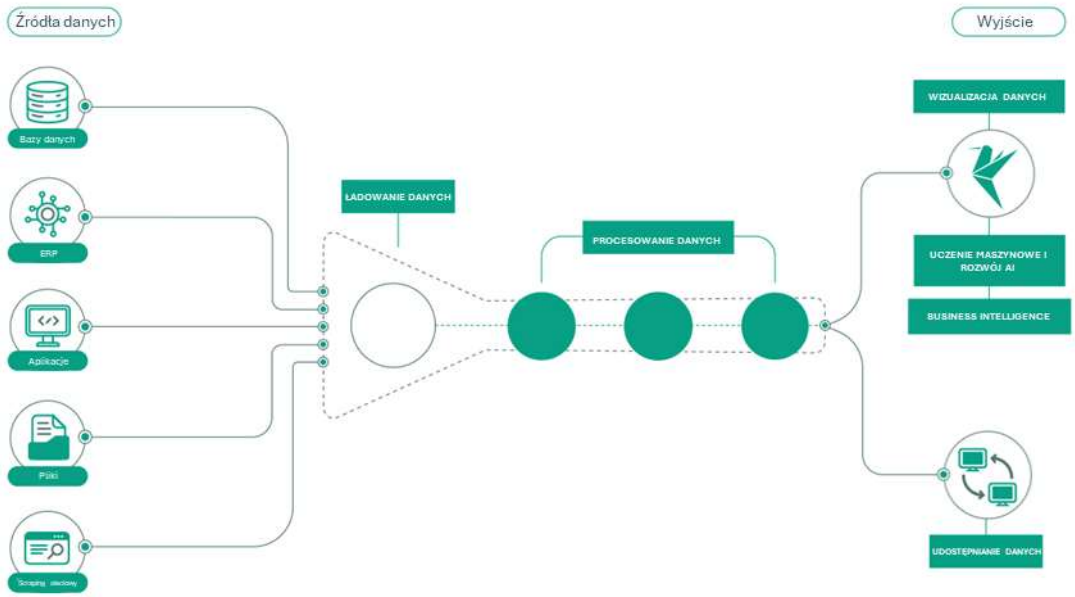


Rysunek 4.3 Etapy analizy danych ML w R

Źródło: Wickham i in. (2023)

W przypadku ML zwykle panuje **błędne przekonanie, że większość czasu i wysiłku poświęca się na faktyczne budowanie algorytmów ML**. W rzeczywistości jest zupełnie odwrotnie, większość czasu poświęca się zazwyczaj na zmagania z danymi i zadaniami wstępnego przetwarzania, a nie na proces modelowania. Czasami wszystkie procesy poprzedzające proces modelowania są o wiele bardziej wymagające i stawiające wyzwania użytkownikom. Dlatego nie ma żadnych przeciwwskazań co do tego, jak te kroki należy wykonać. Rysunek 2.9 przedstawia przykład dobrego podejścia do przekształcania danych w spostrzeżenia biznesowe i wiedzę ogólną. Procedura rozpoczyna się od kroku importu, który jest jednym z najważniejszych kroków w budowaniu modeli ML, ponieważ bez importowania danych do oprogramowania nie można przeprowadzić żadnego rodzaju analizy. Zazwyczaj oznacza to, że należy pobrać dane przechowywane w pliku, bazie danych lub interfejsie programowania aplikacji internetowych (API) i załadować je do ramki danych w R (Wickham i in, 2023). Drugi krok jest związany z uporządkowaniem danych, co jest procedurą unikalną dla R i dotyczy przekształcania danych do określonej formy w celu ich dalszej analizy (każda kolumna jest zmienną, a każdy wiersz jest ramką danych obserwacja-Tibble). Następny krok jest związany z przekształceniem danych, co zwykle obejmuje zawężenie zestawu obserwacji do podpróby zainteresowania. Ponadto może również obejmować tworzenie nowych zmiennych jako kombinacji kilku istniejących zmiennych lub generowanie statystyk podsumowujących. Wizualizacja i modelowanie pełnią odrębne, ale uzupełniające się role

w dziedzinie analizy danych. Wizualizacja jest głęboko zorientowaną na człowieka czynnością, oferującą spostrzeżenia, które mogą umknąć bardziej sformalizowanym podejściom. Dobrze opracowana wizualizacja może ujawnić nieoczekiwane wzorce, skłonić do nowych pytań, a nawet zasugerować, że oryginalne pytania mogą wymagać dopracowania lub innych danych. Natomiast modele zapewniają matematyczne lub obliczeniowe ramy do odpowiadania na precyzyjnie sformułowane pytania. Oferują skalowalność i wydajność, dzięki czemu nadają się do obsługi dużych zestawów danych. Jednakże modele (w których ML jest również uwzględnione) cechują się nieodłącznymi założeniami i nie ma możliwości kwestionowania ani podważania tych założeń. W związku z tym modele mogą nie mieć zdolności do zaskakiwania lub ujawniania nieprzewidzianych spostrzeżeń. Synergia między wizualizacją a modelowaniem jest widoczna w ich wspólnej roli w analizie danych. Wizualizacja pomaga w początkowej eksploracji, zachęcając do formułowania precyzyjnych pytań, podczas gdy modele systematycznie dostarczają odpowiedzi w ramach zdefiniowanych parametrów. Rozpoznanie silnych i słabych stron każdego podejścia jest kluczowe, co prowadzi do bardziej wszechstronnego i świadomego procesu analizy danych. Ostatni krok stanowi komunikację, która jest niezbędna do osiągniecia sukcesu analizy danych, ponieważ jeśli informacje nie zostaną dostarczone decydentowi w odpowiedni i spójny sposób, cała analiza może pójść na marne. Kluczowym elementem w analityce danych są modele ML, bez których nie można by wyciągnąć wniosków na temat procesów biznesowych. Aby rozwiązać konkretne problemy ŁD, architektura ogólnych aplikacji Business Intelligence (przedstawiona na Rysunku 4.2) musi zostać lepiej dostosowana, podobnie jak modele ML. W związku z tym, aby przekształcić sposób działania łańcuchów dostaw poprzez zwiększenie efektywności operacyjnej, usprawnienie podejmowania decyzji i dążenie do realizacji celów korporacyjnych, organizacja **Equilibrium AI** opracowała platformę AI i ML przedstawioną na Rysunku 4.4.



Rysunek 4.4 Proces przetwarzania danych i wiedzy ML dla firmy Equilibrium AI

Źródło: Equilibrium AI (b.d.)

Rysunek przedstawia dobry przykład codziennej praktyki, w jaki sposób ekstrakcja wiedzy i spostrzeżeń jest generowana w aplikacjach ŁD. Zasadniczo proces składa się z operacji back-end i front-end w celu tworzenia wartości (spostrzeżeń biznesowych) dla użytkowników. Proces back-end rozpoczyna się od ekstrakcji danych z różnych źródeł danych, które zwykle znajdują się w ŁD:

- Bazy danych;
- Systemy planowania zasobów przedsiębiorstwa (SAP, Navigator, Microsoft Dynamics itp.);
- Aplikacje (internetowe API);
- Pliki płaskie (csv, xlsx, JSON itp.);
- Sieć, Internet i inne źródła online.

Każde ze źródeł danych ma inną strukturę, protokoły i odpowiednio procedury dotyczące sposobu ekstrakcji i ładowania danych w celu oczyszczenia i wstępnego przetworzenia przed zastosowaniem algorytmów ML. W związku z tym proces ten jest wykonywany za pomocą narzędzi ładujących dane, które posiadają wstępnie zaprogramowany kod służący eksploracji danych pochodzących z różnych źródeł i przekazywania danych wierszowych do nowej bazy danych, która jest ustrukturyzowana

i zorganizowana w celu możliwości zastosowania modeli ML. Przed zastosowaniem modeli ML istnieje jeden dodatkowy krok zwany wstępnym przetwarzaniem danych. W tym kroku dane wierszowe zebrane od firm są sprawdzane pod kątem błędnych danych wejściowych, wartości nielogicznych, prawidłowej struktury danych wejściowych, obserwacji odstających, powtórzeń, NA, NaN itp. Procedura jest kontynuowana poprzez scalanie danych zewnętrznych z danymi firmy. Dane te są zazwyczaj powiązane z czynnikami zewnętrznymi, które mogą potencjalnie wpływać na obserwowany proces biznesowy ŁD, na przykład dane pogodowe, wskaźnik cen konsumpcyjnych, średni dochód w danym regionie, specyficzne cechy demograficzne w danym obszarze, ceny gazu, fale pandemii, komentarze w mediach społecznościowych na temat produktów firmy itp. Jest to bardzo istotne z uwagi na holistyczne zbieranie wszystkich możliwych czynników (wewnętrznych i zewnętrznych), które mogą wpływać na dany proces biznesowy, co zwiększa szansę, że modele ML znajdą właściwy sygnał w danych i będą w stanie wyciągnąć właściwe wnioski i reguły jakie są główne przyczyny, dla których proces biznesowy zachowuje się tak jak zidentyfikowano w ramach obserwacji.

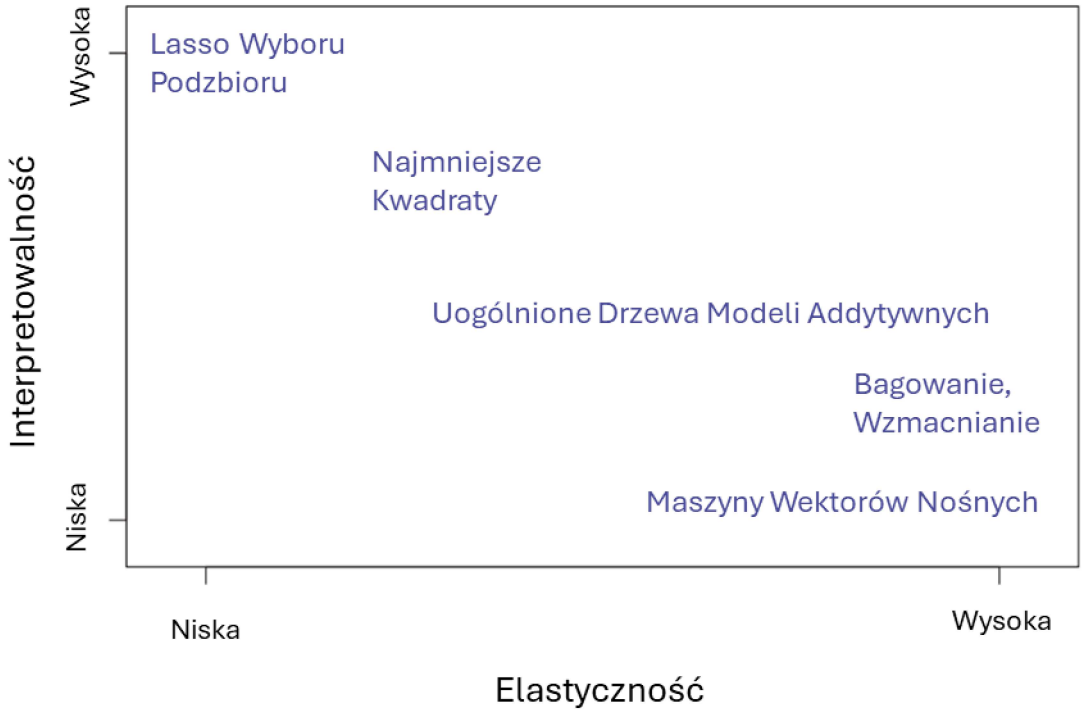
Po połączeniu danych wewnętrznych i zewnętrznych, wstępne przetwarzanie obejmuje wykrywanie sygnału, usuwanie „szumów" z danych, inżynierię cech i losowe dzielenie danych na trening i test (czasem na dane walidacyjne, jeśli opracowano model sieci neuronowej). Dane, które wychodzą z etapu wstępnego przetwarzania, są czyszczone i strukturyzowane w celu możliwości zastosowania modeli ML.

Część wyjściowa składa się z wizualizacji danych, rozwoju ML i AI oraz udostępniania danych. Czasami dane te są udostępniane bez stosowania modeli ML na innych platformach, które przeprowadzają różnego rodzaju analizy (tylko raportowanie do interesariuszy lub agencji rządowych). Proces wizualizacji jest wykonywany za pośrednictwem części front-end platformy, która jest zorientowana na użytkownika i pozwala użytkownikom wykonywać zapytania dotyczące tego, jakie dane łańcucha dostaw, w jaki sposób i w jakich ustawieniach chcą zobaczyć (na przykład Rysunek 4.5).

Rysunek 4.5 Typowa część wizualizacyjna platformy ML w ŁD

Źródło: Opracowanie własne

Z drugiej strony, część przetwarzania danych ML nie jest widoczna dla użytkowników i jest ona niełatwa do zrozumienia. Dlatego modele ML są czasami uważane za **modele czarnej skrzynki**, w ramach których nie ma jasnego zrozumienia, w jaki sposób dokładnie narzędzie połączyło obserwowane dane wejściowe z obserwowanymi danymi wyjściowymi. Jest to jedna z przeszkód, która uniemożliwia szersze wykorzystanie modeli ML w praktyce, zwłaszcza tych, które cechują się dużym stopniem złożoności interpretacji (Rostami-Tabar & Mircetic, 2023). W związku z tym możemy podzielić modele ML na te o wysokiej interpretowalności – niskiej elastyczności i niskiej interpretowalności - wyższej elastyczności (Rysunek 4.6). Ogólnie mówiąc, wraz ze wzrostem elastyczności metody ML, zwykle wzrasta dokładność modelu ML, a jego interpretowalność maleje (Mirčetić i in., 2016).



Rysunek 4.6 Reprezentacja kompromisu (tradeoff) między elastycznością a interpretowalnością przy użyciu różnych metod uczenia maszynowego

Źródło: Hastie i in. (2009)

Dane biznesowe ML i ŁD

Jeśli użytkownicy lepiej rozumieją wizualizację i grafikę, taką jak na Rysunku 4.5, to jaki jest cel wykorzystania ML i czy możliwe jest pominięcie modelowania danych za pomocą algorytmów uczenia maszynowego i zastąpienie ich grafikami informacyjnymi? Niestety nie. Być może głównym powodem, dla którego modele ML są potrzebne, jest fakt, iż nie we wszystkich sytuacjach możliwe jest uzyskanie łatwo czytelnych i wykrywalnych wzorców w danych prezentowanych za pomocą grafiki (jak na Rysunku 4.5). Bardziej powszechną sytuacją jest taka, w której grafika zazwyczaj nie może ujawnić tajemnicy tego, co dzieje się w obserwowanych danych biznesowych ŁD. Dlatego tez potrzebne jest wykorzystanie silniejszych narzędzi w postaci algorytmów ML, celem głębszego wnikania w dane i wyszukiwania **reguł generowania danych** (Rysunek 4.7).

Rysunek 4.7 Charakterystyka statystyczna produktów w łańcuchu dostaw żywności (podsumowanie dla wszystkich produktów)

Źródło: Opracowanie własne

Wyciągniecie prostych wniosków z Rysunku 4.7 i wyprowadzenie reguł biznesowych dotyczących procesu generowania danych jest bardzo trudne. Aby zidentyfikować wzorce zaszyte w danych pochodzących z rysunku, opracowany został model ML, który można wykorzystać do podsumowania cech i wykrycia ważnych sygnałów ukrytych w danych. W związku z tym w Równaniu 1 przedstawiono opracowany model ML dla łańcucha dostaw żywności. Podstawowym czynnikiem napędowym i kręgosłupem tego modelu ML jest autoregresyjny zintegrowany model średniej ruchomej, którego ogólna postać jest następująca:

$$y_t' = c + (\phi_1 y_{t-1}' + \ldots + \phi_p y_{t-p}') + (\theta_1 e_{t-1} + \ldots + \theta_q e_{t-q}) + e_t \text{;} \quad (1)$$

$$y_t - y_{t-1} = c + \phi_1 (y_{t-1} - y_{t-2}) + \ldots + \phi_p (y_{t-p} - y_{t-p-1}) + (\theta_1 B e_t + \ldots + \underbrace{\theta_q e_{t-q}}_{\theta_q B^q e_t} + e_t) \text{;}$$

$$y_t - B y_t = c + \phi_1 (y_{t-1} - B y_{t-1}) + \ldots + \phi_p (y_{t-p} - B y_{t-p}) + (e_t (1 + \theta_1 B + \ldots + \theta_q B^q)) \text{;}$$

$$(1 - B) y_t = c + \phi_1 (1 - B)(y_{t-1}) + \ldots + \phi_p (1 - B) y_{t-p} + e_t (1 + \theta_1 B + \ldots + \theta_q B^q) \text{;}$$

$$(1 - B) y_t = c + \phi_1 (1 - B) B y_t + \ldots + \phi_p (1 - B) B^p y_t + e_t (1 + \theta_1 B + \ldots + \theta_q B^q) \text{;}$$

$$\underbrace{(1 - B)^d y_t}_{\substack{differencing \\ d\_\deg ree}} \cdot \underbrace{(1 - \phi_1 B - \ldots - \phi_p B^p)}_{AR(p)} = c + \underbrace{e_t (1 + \theta_1 B + \ldots + \theta_q B^q)}_{MA(q)} \cdot$$

Model ML wyraźnie pokazuje swoją niską interpretowalność i cechy czarnej skrzynki. Przeciętnemu użytkownikowi biznesowemu trudno jest zrozumieć powiązania między danymi wejściowymi i wyjściowymi. Co więcej, przeciętny użytkownik biznesowy po zapoznaniu z przedstawionym modelem powinien zadać pytanie: Czym jest równanie (1)? Można argumentować, że równanie (1) przedstawia reguły umieszczone na Rysunku 4.1, wygenerowane przez dane ML i potok wiedzy, które ujawniają tajemnicę procesów generowania danych w danym środowisku biznesowym ŁD.

Na pierwszy rzut oka opracowany model ML w równaniu (1) nie wydaje się przyczynić do poprawy zrozumienia danych. Nadal występuje pewna doza dezorientacji, jak w przypadku Rysunku 4.7, jednakże model ML ma kluczową przewagę nad rysunkiem. W istocie model ML jest **wzorem matematycznym**, który może nie być łatwo zrozumiany przez użytkownika, ale jest całkowicie zrozumiały dla narzędzia komputerowego, które **można zaprogramować tak, aby używało dany wzór** i podejmowało **decyzje biznesowe** w oparciu o odkryte reguły.

BIBLIOGRAFIA

1. Chollet, F. (2021). Deep learning with Python. Simon and Schuster.
2. Equilibrium AI (n.d.). Equilibrium AI Data Pipeline [dostępne: https://eqains.com/, dostęp: 23 stycznia 2024]
3. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: Data mining, inference, and prediction (Vol. 2). Springer.
4. Jordan, M. I. & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), s. 255-260.
5. Mirčetić, D., Ralević, N., Nikoličić, S., Maslarić, M. & Stojanović, Đ. (2016). Expert system models for forecasting forklifts engagement in a warehouse loading operation: A case study. Promet-Traffic & Transportation, 28(4), s. 393-401.
6. Ng, A. (2017). Machine learning yearning. [dostępne: http://www.mlyearning.org/, dostęp: 23 stycznia 2024]
7. Rostami-Tabar, B. & Mircetic, D. (2023). Exploring the association between time series features and forecasting by temporal aggregation using machine learning. Neurocomputing, 548, 126376.
8. ScienceSoft (b.d.). Microsoft Business Intelligence to Drive Robust Analytics and Insightful Reporting dostępne: https://www.scnsoft.com/services/business-intelligence/microsoft, dostęp: 23 stycznia 2024]
9. Wickham, H., Çetinkaya-Rundel, M. & Grolemund, G. (2023). R for data science. O'Reilly Media, Inc.