Power BI
i sztuczna inteligencja

Jak w pełni wykorzystać funkcje AI dostępne w Power BI

Mary-Jo Diepeveen

Helion

Tytuł oryginału: Artificial Intelligence with Power BI: Take your data analytics skills to the next level by leveraging the AI capabilities in Power BI

Tłumaczenie: Krzysztof Sawka

ISBN: 978-83-8322-780-1

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/pobisz
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

CZĘŚĆ 2. Gotowe funkcje SI

Wprowadzenie do sztucznej inteligencji w Power BI

Rozdział 1

Każdy chce pracować z danymi. Organizacje wolą podejmować decyzje na podstawie danych niż polegać na intuicji. Aby móc opierać się w procesie decyzyjnym na danych, musimy wyciągać z nich wnioski. Na szczęście **Power BI** jest znakomitym narzędziem do wizualizowania i przekazywania wiedzy zawartej w danych. Aby lepiej zrozumieć, jakie trendy i wnioski możemy odczytać z danych, możemy się posłużyć technikami wykorzystywanymi w danetyce.

Danetyka (ang. *data science*) i **sztuczna inteligencja** (**SI**) stopniowo stają się coraz popularniejszymi technikami wydobywania wniosków z danych. Wynika to między innymi z faktu, że narzędzia te umożliwiają pracę na nieustrukturyzowanych danych, co wcześniej nie było możliwe. Pozwala to szybciej wyszukiwać skomplikowane trendy i wzorce w danych.

W tej książce skoncentrujemy się na korzystaniu z aplikacji Microsoft Power BI jako narzędziu eksplorowania i wizualizowania danych. Skorzystamy też z pewnych elementów **platformy rozproszonej Azure** umożliwiających trenowanie modeli oraz ich integrację z Power BI.

Najpierw jednak zajmijmy się podstawami. Musimy zrozumieć, czym jest SI, aby umieszczać projekty w odpowiednich ramach i skutecznie je realizować. Musimy wiedzieć, co jest możliwe i jak przejść od prostego modelu do modelu SI, zanim zajmiemy się szczegółami każdego etapu tego procesu. Dlatego najpierw odpowiemy sobie na następujące pytania:

- Czego możemy oczekiwać od analityka danych?
- Czym jest sztuczna inteligencja?
- Dlaczego powinniśmy używać SI w Power BI?
- Jakimi rodzajami SI dysponujemy w Power BI?

Zacznijmy od tych pytań.

Czego możemy oczekiwać od analityka danych?

Każdą firmę interesują inne spostrzeżenia płynące z danych i pracuje na innych rodzajach i zestawach danych. Nawet jeżeli w każdej organizacji znajdziesz analityka danych, zakres obowiązków osób na tym stanowisku może się diametralnie różnić. Podczas lektury tej książki skoncentrujesz się na elementach przydatnych dla Ciebie i prawdopodobnie pominiesz nieistotne fragmenty. Mimo to dobrze jest wiedzieć, jakiej wiedzy oczekujemy od Ciebie oraz z czym powinieneś być zaznajomiony.

Najpierw wyjaśnimy, co to znaczy być analitykiem danych, i omówimy przyjmowane założenia i powód wybrania takiej nazwy profesji. Następnie przejdziemy do tego, co już zapewne wiesz o Power BI, oraz podamy źródła informacji, aby umożliwić Ci odświeżenie tej wiedzy.

Kim jest analityk danych?

Możesz nazwać siebie inżynierem analityki biznesowej, specjalistą ds. analityki biznesowej, administratorem bazodanowym lub po prostu analitykiem danych. Bez względu na nazwę swojej profesji wybrałeś tę książkę, ponieważ pracujesz w Power BI i chcesz dowiedzieć się więcej o tym pakiecie. Obecnie, gdy na rynku istnieje takie zatrzęsienie nazw profesji, coraz trudniej ustalić podstawową wymaganą wiedzę. Dla zachowania prostoty i spójności będziemy nazywać osobę pracującą w Power BI analitykiem danych.

Dlaczego analitykiem danych? Ponieważ zakładamy w tej książce, że potrafisz pracować na danych w Power BI oraz wykonywać następujące czynności:

- przygotowywać dane;
- modelować dane (tworzyć model danych w Power BI, a nie model uczenia maszynowego);
- wizualizować dane;
- analizować dane;
- wdrażać i konserwować elementy Power Bi.

Z drugiej strony założymy, że nic nie wiesz na temat danetyki. Przyjrzymy się wszystkim cechom SI w Power BI z perspektywy takiej osoby z nadzieją, że nauczysz się korzystać z nich we właściwy sposób. Nie będziemy jednak wchodzić w szczegóły każdego modelu,

gdyż książka ta nie jest napisana z myślą o danetykach, którzy mają już rozległą wiedzę o metodach matematycznych i statystycznych używanych w SI.

Podczas podążania ścieżką SI w Power BI główną rolę odgrywają dwie cechy: łączenie się z danymi oraz wizualizowanie danych. Przyjrzyjmy się dokładniej tym cechom, dzięki czemu będziesz wiedzieć, czego się spodziewać, zanim przejdziemy dalej.

Łączenie się z danymi

Skoro jesteśmy danetykami, przyjrzyjmy się, jakie są nasze główne zadania. Pierwszą czynnością podczas pracy z danymi jest **uzyskiwanie dostępu do danych**. Z perspektywy technicznej możemy bardzo łatwo połączyć Power BI z różnymi źródłami danych bez względu na to, czy dane są przechowywane w rozproszonych bazach danych (np. Azure), czy mają postać plików lokalnych. Power BI pozwala łączyć się z danymi w każdym przypadku, a nawet umożliwia harmonogramowanie automatycznego odświeżania w celu wizualizowania nowych danych tak długo, jak będziemy utrzymywać połączenie między siecią, w której przechowywane są dane, a usługą Power BI.

Na jakiego typu danych możemy pracować? Na dowolnych danych! Podczas wprowadzania obrazów do raportów Power BI możesz łączyć się z danymi **ustrukturyzowanymi**, czyli sformatowanymi w ładne tabele, **częściowo ustrukturyzowanymi** (często występującymi w formacie JSON), a nawet **nieustrukturyzowanymi**. Oznacza to także, że dane mogą pochodzić z różnych źródeł. Możesz gromadzić dane Twittera (częściowo ustrukturyzowane), zawierające tekst wpisu, datę jego utworzenia, liczbę udostępnień, polubień czy hasztagów. Możesz zbierać dane marketingowe i sprzedażowe, aby dowiedzieć się, które produkty sprzedałeś, kiedy je sprzedałeś, a także które prowadzone przez Ciebie kampanie mogłyby mieć wpływ na sprzedaż. A może sprawdzasz podaż i popyt dla swoich produktów, gdyż chcesz odpowiednio zaplanować logistykę w magazynach i sklepach.

Skoro dane są generowane przez tak wiele różnych źródeł i mogą występować w wielu różnych formatach, chcemy również wiedzieć, jak *wydobyć te dane* i przygotować do raportów. Power BI zawiera liczne standardowe łączniki umożliwiające łączenie z danymi. Najlepszym rozwiązaniem tutaj jest jednak opracowanie potoku zarządzającego obsługą danych jeszcze przed ich podłączeniem do Power BI. Proces taki jest nazywany potokiem **ETL** (ang. *Extract-Transform-Load* — wydobądź – przekształć – załaduj) lub **ELT** (ang. *Extract-Load-Transform* — wydobądź – załaduj – przekształć), łączącym się ze źródłami generującymi dane, wydobywającym te dane, ładującym je do bazy danych i w razie potrzeby przekształcającym je. Mimo że podobne zadania można realizować w Power BI, wolimy korzystać z narzędzi ETL takich jak Azure Data Factory, obsługujących tego typu potoki podczas pracy na znacznych ilościach danych.

ETL czy ELT?

ETL jest ugruntowanym i powszechnie stosowanym sposobem wydobywania danych ze źródeł. Jego celem jest często przekształcanie danych i ich umieszczanie w ustrukturyzowanej bazie danych, takiej jak **Azure SQL Database**, oraz dopasowywanie ich do kolumn i wierszy. Jest to świetne rozwiązanie w przypadku danych transakcyjnych, gdzie chcemy uzyskiwać szybko rezultaty. Jednak wraz z nastaniem ery przetwarzania rozproszonego coraz większą uwagę zyskuje nowsze podejście, czyli ELT. Poprzez wydobywanie danych i umieszczanie ich w pamięci masowej mogącej przechowywać dane nieustrukturyzowane, takiej jak **Azure Storage Account** czy **Azure Data Lake**, możemy posiadać dane w chmurze bez konieczności troszczenia się o dalszy plan. W ten sposób możliwe jest również wielokrotne wykorzystywanie tych samych danych oraz ich przekształcanie na różne sposoby, w zależności od wniosków, jakie chcesz z nich wydobyć.

Mówiąc krótko, z danymi może wiele się dziać (i prawdopodobnie się dzieje), jeszcze zanim w ogóle otworzysz Power BI. Miej świadomość procesów, jakie mogły mieć miejsce jeszcze przed wprowadzeniem danych do Power BI. Każde źródło, które podamy aplikacji Power BI, będzie wpływać na dostępne w niej opcje, jak również na wydajność generowanych raportów. W tej książce będziemy pracować głównie na już wstępnie przetworzonych danych, dostępnych w plikach umieszczonych w publicznych serwisach lub rozproszonych bazach danych. Jednak w przypadku niektórych projektów angażujących SI będziemy musieli przetwarzać dane przed wprowadzeniem ich do Power BI, aby można było na nich pracować.

Wizualizowanie danych

Ostatnim powodem, dla którego będziemy korzystać z Power BI, jest fakt, że *za pomocą danych chcemy opowiedzieć historię.* Chcemy przekształcić dane w cenne i intuicyjne spostrzeżenia, które każdy członek naszej organizacji będzie mógł analizować na swój własny użytek. W kontekście sztucznej inteligencji może to być jedna z najważniejszych umiejętności analityka danych, co zostanie wyjaśnione w dalszej części rozdziału.

Co więc rozumiemy przez opowiadanie historii? Ludzie są znacznie bardziej podatni na słuchanie opowieści niż suchych danych. Jeżeli powiemy, że na 70% będzie padać deszcz, to czy weźmiesz parasolkę? Trudno powiedzieć. Jeżeli ktoś Ci powie, żebyś wziął parasolkę, to prawdopodobnie to zrobisz. Nawet jeżeli chcesz częściej podejmować decyzje na podstawie danych, ludzie z natury nie są do tego skłonni. Oddziałują na nas opowieści, które są bardziej intuicyjne. Oznacza to także, że nie możemy przekazywać człowiekowi ot tak spostrzeżeń uzyskiwanych przez SI. Musimy *przetłumaczyć* wynik modelu SI w taki sposób, żeby stał się zrozumiały dla ludzi. Oznacza to, że musimy za pomocą danych opowiadać historie.

Możemy to zrobić w Power BI dzięki dostępnym możliwościom wizualizacji. Możemy korzystać ze standardowych wizualizacji, importować wizualizacje ze sklepu albo tworzyć własne wizualizacje za pomocą języków **Python** lub **R**. Istotną cechą analityka danych (którą, jak się domyślamy, dysponujesz podczas czytania tej książki) jest świadomość tego, kiedy używać odpowiednich narzędzi oraz jak łączyć różne wizualizacje w raport. Podczas lektury nie zapominaj, że aby zdobyć zaufanie ludzi, należy mówić ich językiem, a nie podawać im suche liczby z nadzieją, że wpłyną one na ich zachowanie.

W tej książce będziemy koncentrować się na sposobach wykorzystania Power BI w połączeniu ze sztuczną inteligencją. Oznacza to, że Ty jako analityk danych znasz już różne typy danych, które możesz wprowadzać do Power BI z różnych dostępnych źródeł. Powinieneś umieć już przygotowywać wizualizacje tworzące raporty w Power BI. W kolejnych rozdziałach będziemy kłaść nacisk na funkcje mające znaczenie podczas przygotowywania danych z myślą o SI oraz podczas implementowania tejże sztucznej inteligencji. Najpierw jednak wyjaśnijmy sobie, czym właściwie jest sztuczna inteligencja.

Czym jest sztuczna inteligencja?

Sztuczna inteligencja jest terminem stosowanym często przez organizacje pragnące podkreślać, że korzystają z najnowszych zdobyczy technologii. Co ciekawe, pojęcie jest już znane od ponad 60 lat. Na początku definiowano ją jako *naukę i inżynierię tworzenia inteligentnych maszyn* (profesor John McCarthy, Uniwersytet Stanforda, *http://jmc.stanford.edu/artificial-intelligence/what-is-ai/*, dostęp w lutym 2023 r.). Niestety dla nas pozostawia to mnóstwo miejsca na interpretacje, między innymi przez co termin ***SI*** zyskiwał z upływem lat tak wiele różnych znaczeń.

SI idzie często w parze z danetyką, stanowiącą inżynierii, matematyki i statystyki. Jej zadaniem jest wydobywanie spostrzeżeń z danych oraz wyjaśnianie surowych danych pozyskiwanych z dostępnych nam aplikacji bądź baz danych. Dzięki danetyce możemy zdobywać dane, oczyszczać je, trenować za ich pomocą model, który następnie można zintegrować z naszymi aplikacjami, aby uzyskiwać przewidywania na podstawie nowych danych.

Aby w pełni zrozumieć możliwości sztucznej inteligencji, musimy poznać kilka różnych pojęć, często stosowanych w powiązaniu z SI: **uczenie maszynowe**, **uczenie głębokie**, **uczenie nadzorowane** i **uczenie nienadzorowane**. Pomogą nam one również w zaznajomieniu się z typową strukturą procesu budowania modelu SI.

Definicja sztucznej inteligencji

Istnieje wiele różnych definicji sztucznej inteligencji, w których zasadniczo występują trzy wspólne elementy:

- komputery
- wykonujące inteligentne zadanie
- w taki sposób, w jaki wykonałby je człowiek.

Komputery występują w najróżniejszych postaciach i mogą oznaczać zarówno oprogramowanie, jak i sprzęt; może to być aplikacja działająca lokalnie na czyimś laptopie lub najprawdziwszy robot. Bardziej otwartą na interpretacje częścią definicji SI jest realizacja inteligentnego zadania w taki sposób, w jaki wykonałby je człowiek. Co oznacza tu słowo *inteligentne*? Jeżeli pomyślimy o inteligentnym zadaniu wykonywanym przez człowieka, możemy posłużyć się przykładem kalkulatora. Droższe kalkulatory mogą w ciągu kilku sekund wykonać skomplikowane obliczenia, co matematykowi zajęłoby znacznie więcej czasu. Gdybyśmy jednak zapytali kogoś, czy można uznać kalkulator za sztuczną inteligencję, to prawdopodobnie usłyszelibyśmy głośne *nie*.

Rodzi się więc pytanie: czym jest *inteligencja*? Na szczęście wielu filozofów poświęca karierę akademicką poszukiwaniom odpowiedzi na to pytanie, dlatego przyjmijmy, że wykracza ona poza ramy tematyczne tej książki. Przyjmijmy, że granica tego, co jest uznawane za sztuczną inteligencję, przesuwa się z roku na rok. Z kolejnymi postępami rodzą się nowe oczekiwania. Kiedyś uważaliśmy pokonanie mistrza świata w szachach za szczytowe osiągnięcie SI, obecnie zaś poważnie zastanawiamy się nad możliwością stworzenia w pełni autonomicznych samochodów.

Wśród nowych wynalazków pojawiły się algorytmy umożliwiające uczenie jeszcze inteligentniejszych modeli. Algorytmy są często określane jako należące do dziedziny **uczenia maszynowego** (ang. *machine learning*) oraz **uczenia głębokiego** (ang. *deep learning*); jest bardzo ważne, aby rozróżniać zakres zadań realizowanych w ramach każdej z tych dziedzin. Omówimy teraz szczegółowo każde z tych pojęć.

Definicja uczenia maszynowego

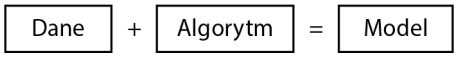
Jeśli wrócimy do przykładu prostego, ale wydajnego kalkulatora, to możemy wyobrazić sobie inteligencję takiej maszyny jako opracowaną w systemie reguł. Gdy dodamy 1 do 1, zawsze otrzymamy 2. Takie (i wiele innych) reguły matematyczne możemy tak zaprogramować w kalkulatorze, aby był w stanie liczyć. Technika taka jest nazywana **stosowaniem wyrażeń regularnych** i do dzisiaj okazuje się bardzo przydatna. Uznawana jest za najbardziej ograniczoną metodę otrzymywania SI, ale uzyskuje szybkie i zrozumiałe wyniki.

Jeśli jednak liczysz na mądrzejszą SI, musisz zainteresować się technikami, w których model nie jest w pełni zaprogramowany zgodnie z wyznaczonymi przez człowieka regułami, lecz ma możliwość **samouczenia**. Z powodu tego pojęcia często uważa się SI za **samodoskonalący się** twór, który będzie wciąż się rozwijał aż do osiągnięcia punktu **osobliwości**. W rzeczywistości jednak samouczenie oznacza, że nie musimy otwarcie mówić sztucznej inteligencji, jak ma interpretować dostarczane dane. Pokazujemy jej natomiast mnóstwo przykładów, na których podstawie trenowany przez nas model będzie decydował, w jaki sposób wzorzec zmiennych wartości wpływa na określone przewidywania.

Na przykład możesz sprzedawać laptopy i chcieć zareklamować odpowiedni model odpowiedniej osobie. Możesz to zrobić za pomocą systemu reguł, w którym stworzysz grupy na podstawie danych demograficznych, np. kobiety przed trzydziestką i kobiety po trzydziestce oraz taki sam podział mężczyzn. Mielibyśmy cztery różne grupy, wobec których moglibyśmy stosować różne strategie marketingowe, przy założeniu, że każda kobieta przed trzydziestką ma takie same wymogi co do nowego laptopa itd.

Oczywiście chcemy zamiast tego skorzystać ze wzorców, których sami nie dostrzegamy, ale które znajdują się w danych. Właśnie w takiej sytuacji użylibyśmy uczenia maszynowego do stworzenia modelu samouczącego, który przeglądałby dane i wyszukiwałby zmienne lub cechy szczególnie interesujące w kwestii wymagań klienta. Na podstawie tego samouczenia mogłoby się okazać, że istnieją zupełnie inne grupy, wobec których należałoby stosować odmienne strategie marketingowe. Na przykład mogłaby zostać wyznaczona grupa kobiet i mężczyzn przed trzydziestką uwielbiających gry sieciowe, której wymagania będą naturalnie odmienne od wymagań grupy kobiet i mężczyzn przed trzydziestką używających laptopa wyłącznie do pracy.

W porównaniu do wyrażeń regularnych korzystanie z uczenia maszynowego do uzyskania SI jest bardziej skomplikowanym zadaniem. Aby stworzyć model uczenia maszynowego, bierzemy **dane** i wybieramy **algorytm** służący do **trenowania modelu**. Pojęcia te zostały zwizualizowane na rysunku 1.1.



Rysunek 1.1. Tworzenie modelu uczenia maszynowego

Jak widać na rysunku 1.1, posiadane przez nas dane i wybrany algorytm są wejściami, natomiast wytrenowany model jest wyjściem. Wybrany przez nas algorytm określa sposób traktowania danych. Czy chcemy je klasyfikować? A może chcemy stworzyć model regresyjny, w którym będzie przewidywana wartość numeryczna? Albo może chcemy pogrupować dane w oddzielne skupienia? Informacje te są zawarte w algorytmie wybranym do trenowania modelu.

Skoro już wiemy, czym jest uczenie maszynowe, dowiedzmy się, w jaki sposób różni się ono od uczenia głębokiego.

Definicja uczenia głębokiego

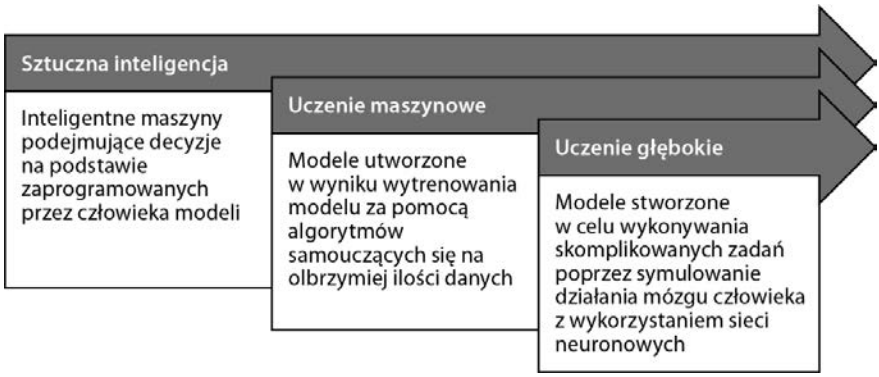
Już uczenie maszynowe samo w sobie otworzyło danetykom furtkę do całego świata nowych możliwości. Zamiast poświęcać niezliczone godziny na eksplorowanie danych oraz liczenie **korelacji**, **kowariancji** i innych wskaźników statystycznych w celu znajdowania wzorców w danych, wystarczy wytrenować model do wyszukiwania ich za nas.

Uczenie maszynowe było początkowo realizowane na ustrukturyzowanych danych umieszczonych elegancko w kolumnach i wierszach. Wkrótce jednak przestało to wystarczyć. Danetycy pragnęli także klasyfikować obrazy oraz rozpoznawać wzorce w olbrzymich dokumentach tekstowych. Niestety algorytmy uczenia maszynowego nie radzą sobie z takimi nieustrukturyzowanymi danymi, głównie z powodu poziomu skomplikowania tychże danych. Mówi się, że obraz jest wart tysiąca słów, i rzeczywiście, nawet pojedynczy piksel obrazu przechowuje mnóstwo informacji, które można przeanalizować na wiele różnych sposobów.

Wraz z nastaniem **obliczeń rozproszonych** i rozwojem procesorów pojawiła się nowa dziedzina uczenia maszynowego. Obecnie dysponujemy już nie tylko prostymi **procesorami głównymi** (ang. *central processing unit* — **CPU**), ale również potężniejszymi **procesorami graficznymi** (ang. *graphical processing unit* — **GPU**), które znacznie szybciej przetwarzają skomplikowane dane, takie jak obrazy. Większa moc oznacza wyższe koszty, ale dzięki usługom rozproszonym mamy do dyspozycji jednostki GPU na żądanie i płacimy wyłącznie wtedy, gdy z nich korzystamy.

Pomimo dostępności większej mocy obliczeniowej wciąż nam brakowało nowych algorytmów będących w stanie wydobywać wzorce z takich nieustrukturyzowanych danych. Chcieliśmy realizować te zadania tak samo, jak robiłby to człowiek, dlatego naukowcy zainteresowali się ludzkim mózgiem i przyjrzeli się jego mechanizmom przetwarzania informacji. Mózg składa się z komórek zwanych **neuronami**, połączonych w wiele warstw, przez które przechodzą różnorakie sygnały. Z tego powodu badacze spróbowali odtworzyć uproszczoną **sztuczną sieć neuronową**, w której symulowane są te neurony i warstwy. Wyniki okazały się bardzo dobre i doprowadziły do powstania dziedziny uczenia głębokiego. Teraz możemy klasyfikować obrazy lub wykrywać znajdujące się na nich obiekty za pomocą **widzenia komputerowego** (**maszynowego**; ang. *computer vision*). Z kolei wydobywaniem spostrzeżeń z danych tekstowych zajmuje się **przetwarzanie języka naturalnego** (ang. *natural language processing* — **NLP**).

Dowiedzieliśmy się już co nieco na temat sztucznej inteligencji, uczenia maszynowego i uczenia głębokiego. Zależności między tymi trzema pojęciami zostały zaprezentowane na rysunku 1.2.



Rysunek 1.2. Związek między sztuczną inteligencją, uczeniem maszynowym i uczeniem głębokim

Jak widać na rysunku 1.2, te trzy pojęcia są ze sobą ściśle powiązane. SI jest często uznawana za najbardziej ogólne pojęcie, w którym mieści się wszystko, co robimy, aby umożliwić maszynie realizowanie inteligentnego zadania tak, jak wykonuje je człowiek. Jedną z obranych strategii stało się uczenie maszynowe, w którym trenujemy modele, nie korzystając z reguł, lecz umożliwiając im samouczenie się. My jedynie dostarczamy dane i algorytm (definiujący zadanie) do uczenia modelu. Z kolei uczenie głębokie jest dziedziną uczenia maszynowego, w której wyspecjalizowane algorytmy służą do lepszego realizowania skomplikowanych zadań takich jak wyszukiwanie wzorców w nieustrukturyzowanych danych.

Poza tymi trzema terminami, które należy zrozumieć przed rozpoczęciem pracy ze sztuczną inteligencją w Power BI, musimy jeszcze pojąć różnicę między **uczeniem nadzorowanym** (ang. *supervised learning*) a **uczeniem nienadzorowanym** (ang. *unsupervised learning*). Obydwie techniki dzielą wykorzystywane przez nas algorytmy na dwie kategorie. Zrozumienie występujących między nimi różnic pomoże określić, co należy uwzględnić w zestawach danych jako dane wejściowe dla trenowanego modelu.

Uczenie nadzorowane a uczenie nienadzorowane

Zadaniem SI jest uzyskiwanie przewidywań. Chcesz przewidywać na przykład takie rzeczy jak to, czy ktoś prędzej kupi zmywarkę, czy nową lodówkę. A może wolisz przewidzieć, ile danego dnia sprzedasz jabłek, dzięki czemu będziesz wiedzieć, jak zaopatrzyć sklep. Przewidywany element jest często nazywany **etykietą** (ang. *label*) lub **znacznikiem** (ang. *tag*). Czasami dysponujemy danymi uczącymi zawierającymi taką etykietę, a czasem nie.

Jeśli prowadzimy sklep zajmujący się sprzedażą jabłek, możemy połączyć dane historyczne na temat sprzedaży z dodatkowymi danymi, takimi jak pogoda i dzień

tygodnia. W zimne poniedziałki nikt nie chce kupować jabłek, ale w słoneczne piątki może zabraknąć jabłek na półkach jeszcze przed południem. Skoro możemy sprawdzić w danych historycznych, ile jabłek sprzedawaliśmy w przeszłości w określonych warunkach, to dysponujemy danymi uczącymi zawierającymi etykietę, mianowicie liczbę sprzedanych jabłek. Jeżeli znamy etykietę, to mamy do czynienia z **uczeniem nadzorowanym**.

A co w przypadku sprzedaży laptopów? Powiedzmy, że dysponujemy danymi klientów zawierającymi informacje demograficzne, takie jak wiek czy płeć. Możemy jednak mieć także dane dotyczące celu użytkowania laptopów: do gier sieciowych lub do pracy. W takim przypadku nie wiemy, dla ilu grup należałoby utworzyć różne strategie marketingowe. Zatem w danych uczących nie istnieją grupy, w jakie chcemy skategoryzować klientów. Nie ma więc w tych danych etykiet, przez co mamy do czynienia z **uczeniem nienadzorowanym**.

Dobrze jest odróżniać te dwa pojęcia, ponieważ łatwiej będzie Ci określać, czego należy oczekiwać od danych oraz od modelu. Na koniec przyjrzymy się różnym rodzajom algorytmów, z których będziemy korzystać w tej książce.

Rodzaje algorytmów

Gdy postanowiliśmy zastosować SI do swoich danych, to wiemy już, że potrzebujemy danych oraz algorytmu do utworzenia modelu. W kolejnych rozdziałach omówimy znacznie dokładniej wymogi dotyczące danych. Pozostaje nam więc kwestia algorytmów. Sposób pracy z algorytmami jest postrzegany jako wchodzący w zakres kompetencji danetyków, gdyż łączą się w nim dziedziny matematyki i statystyki. Jednak nawet jeśli sami nie budujemy modeli lub nie chcemy być pełnoetatowymi danetykami, zawsze warto zrozumieć główne rodzaje algorytmów, na których przyjdzie nam pracować.

Najważniejsza jest informacja, że wraz z doborem algorytmu określamy sposób postrzegania danych oraz rodzaj wzorca wykrywanego przez model. Jeśli chodzi o uczenie nadzorowane (gdzie znamy przewidywaną etykietę), często mówimy o **regresji** (ang. *regression*) lub **klasyfikacji** (ang. *classification*). W przypadku regresji próbujemy przewidzieć wartość numeryczną, podczas gdy w klasyfikacji etykieta jest kategorialna (zawiera co najmniej dwie kategorie).

Algorytmy regresyjne

Wyobraź sobie, że pracujesz dla krajowego dostawcy prądu. Będziesz mieć zgromadzonych mnóstwo danych na temat klientów, w tym takich jak miejsce zamieszkania, rodzaj i wielkość lokalu czy liczba osób przypadających na gospodarstwo domowe. W przypadku dotychczasowych klientów znasz ich zużycie energii w poprzednich latach. Chcesz przewidzieć zużycie prądu przez nowych klientów, aby być w stanie wiarygodnie oszacować im koszty.

Nasze dane mogą wyglądać jak tabela zaprezentowana na rysunku 1.3, gdzie widzimy dane historyczne dla dwóch dotychczasowych klientów, a chcemy przewidzieć zużycie prądu w kilowatogodzinach dla trzeciego, nowego klienta:

Lokalizacja	Rodzaj lokalu	Liczba osób	Wielkość lokalu (m2)	Zużycie prądu (kWh)
Duże miasto	Mieszkanie	2	50	1990
Małe miasto	Dom	4	120	4320
Duże miasto	Mieszkanie	3	70	???

Rysunek 1.3. Dane gospodarstw domowych trzech klientów wraz z zużyciem energii

Znamy w tym przykładzie etykietę: zużycie prądu. Wiemy więc, że mamy do czynienia z uczeniem nadzorowanym. Zmienna, którą chcemy przewidywać, jest *wartością numeryczną* (np. 1990 kWh, 4320 kWh, a także każda wartość pośrednia). Jest to zatem przykład prostego modelu regresyjnego. Z podzbioru regresyjnego moglibyśmy wybrać różne algorytmy do uczenia tego modelu. Wybór zależy od takich czynników, jak poziom skomplikowania tworzonego modelu i stopień wyjaśnialności modelu, a także od dostępnej mocy obliczeniowej i od czasu obliczeniowego na trenowanie modelu. Do tej kategorii algorytmów należą: **regresja liniowa**, **regresja metodą drzew decyzyjnych** i **regresja metodą wzmocnionych drzew decyzyjnych**.

Algorytmy klasyfikacyjne

Po wytrenowaniu modelu za pomocą jednego z tych algorytmów i danych historycznych byliśmy w stanie właściwie przewidywać zużycie prądu przez nowego klienta oraz koszt tego zużycia. Klient ten zgodził się kupować od nas prąd, ale chce również dowiedzieć się więcej na temat potencjalnych sposobów zaoszczędzenia energii. Zostaje tu poruszony wątek korzystania z paneli fotowoltaicznych. Jeżeli ludzie mają zamontowane panele na dachach domów, mogą generować własną energię w słoneczne dni i zaoszczędzić pieniądze. My chcemy oczywiście pomagać im w takiej inwestycji i w instalacji paneli.

Niektórzy klienci mogą już mieć zamontowane panele fotowoltaiczne, inni mogą chcieć się dowiedzieć więcej na ich temat przed zakupem, a jeszcze innym mogła nie powstać w głowie myśl o ich zainstalowaniu. Chcemy skontaktować się z klientami i zareklamować im sprzedawane przez nas panele, ale nie chcemy irytować klientów ani wydawać budżetu marketingowego na klientów już korzystających z tego rozwiązania.

Zatem chcemy teraz uzyskać katalog klientów będących już właścicielami paneli fotowoltaicznych. Oczywiście moglibyśmy skontaktować się z każdym gospodarstwem domowym, wydaje się to jednak zbyt dużym zadaniem wymagającym zbyt wielkiego

nakładu czasu i energii. Poza tym nie wszystkie gospodarstwa mogłyby wziąć udział w ankiecie mającej na celu zdobycie tych danych. Dlatego postanawiamy sprawdzić, czy uda nam się przewidzieć klientów posiadających panele fotowoltaiczne. Możemy zebrać próbę danych, wytrenować na nich model i użyć uzyskanych spostrzeżeń do wskazania odpowiednich gospodarstw domowych.

Ta próba danych może wyglądać tak jak tabela zaprezentowana na rysunku 1.4. Tutaj znów dysponujemy znanymi/historycznymi danymi oraz etykietą. Skoro znamy etykietę próby danych, to będziemy przeprowadzać uczenie nadzorowane. W tym przypadku jednak nie próbujemy przewidywać wartości numerycznej. Etykieta, którą próbujemy przewidzieć, to *Panele fotowoltaiczne*, które mogą przyjmować wartości *Tak* lub *Nie*. Mamy więc tu dwie kategorie, co oznacza problem klasyfikacyjny, a dokładniej mówiąc **binarny** lub **dwuklasowy** problem klasyfikacji:

Położenie	Rodzaj lokalu	Liczba osób	Wielkość lokalu (m2)	Zużycie prądu (kWh)	Panele fotowoltaiczne
Duże miasto	Mieszkanie	2	50	1990	Nie
Małe miasto	Dom	4	120	4320	Tak
Duże miasto	Mieszkanie	3	70	2490	???

Rysunek 1.4. Parametry gospodarstw domowych i etykieta paneli fotowoltaicznych

Również w tym przypadku są dostępne różne algorytmy, które możemy wybrać, gdy wiemy, że chcemy realizować zadanie klasyfikacji. Przede wszystkim ważna jest liczba klas etykiety. Także w tym przypadku możemy określić złożoność i wyjaśnialność modelu dzięki takim algorytmom uczenia modelu jak **dwuklasowa regresja logistyczna** czy **las losowy**.

Na koniec spójrzmy na jeszcze jeden prosty przykład ukazujący różne rodzaje algorytmów, z którymi przyjdzie nam pracować. Wyobraź sobie, że wciąż pracujemy dla krajowego dostawcy prądu, ale na rynek weszła konkurencja. Obawiamy się, że ta nowa firma odbierze nam część klientów, dlatego chcemy zaoferować klientom jakieś korzyści, by ich zatrzymać. Aby zaoszczędzić pieniądze, nie chcemy ich oferować wszystkim, co oznacza, że musimy odpowiednio ocenić, kto może z nas zrezygnować.

Odeszło od nas na razie względnie niewielu klientów i chcemy, żeby tak pozostało. Oznacza to jednak, że nie mamy etykiety; brakuje nam danych w przewidywanej zmiennej. W tym przypadku możemy pogrupować klientów w skupienia: na tych, którzy przejdą do konkurencji, oraz na tych, którzy pozostaną u nas. W danych mogą

występować jeszcze klienci, którzy już zrezygnowali z naszych usług. Na podstawie tej odrobiny posiadanych informacji możemy skorzystać na przykład z **algorytmu centroidów**, aby wyszukiwać punkty danych podobne do reprezentujących klientów, którzy nas zostawili, utworzyć za ich pomocą wspomniane grupy, a następnie zaproponować korzystne warunki dla tych, którzy najprawdopodobniej nas opuszczą, aby ich zatrzymać.

Praca z algorytmami wymaga znajomości kryjących się za nimi matematyki i statystyki. Aby móc je wykorzystać do trenowania modeli, polegamy na tej wiedzy dostarczanej przez danetyków zatrudnionych w zespole, dzięki czemu jesteśmy w stanie lepiej podejmować decyzje na podstawie danych. Nie musimy być ekspertami danetyki, żeby móc korzystać z funkcji SI w Power BI. Przydaje się jednak rozumienie podejmowanych wyborów, aby mieć świadomość potencjału i ograniczeń korzystania ze sztucznej inteligencji na naszych danych.

Skoro wiemy już, czym jest SI i jak wykorzystywać różne algorytmy do trenowania modelu, wykonajmy krok wstecz i przyjrzyjmy się pełnemu **procesowi danetycznemu**. W jaki sposób traktujemy dane od początku do końca? I jak możemy robić to skutecznie?

Czym jest proces danetyczny?

Uczenie modelu, jak każdy proces, jest wieloetapowe. I podobnie jak w wielu projektach, fazy te *niekoniecznie muszą występować liniowo*. Podczas przygotowywania rozwiązania SI zależy nam raczej na podejściu iteracyjnym. Sprawdźmy najpierw, z jakich faz składa się proces danetyczny.

Przede wszystkim musimy się zastanowić, po co to robimy. Dlaczego chcemy skorzystać z SI? Co ten model będzie robił? Nawet jeżeli motywuje to do innowacyjności, powinniśmy unikać korzystania ze sztucznej inteligencji, jeśli jedynym powodem jest to, że wszyscy z niej korzystają. Mimo to określenie odpowiedniego powodu korzystania ze sztucznej inteligencji bywa czasami trudne, gdyż wiele zastosowań jest względnie nowych i nieznanych. Jaki jest więc dobry powód? To zależy oczywiście od dziedziny, ale w każdym obszarze działalności istnieje bardzo korzystny sektor, który możemy rozpoznać. Najczęściej myślimy o używaniu SI do przewidywania sposobów inwestowania budżetów marketingowych, zwiększania obrotów, monitorowania w kontekście konserwacji predykcyjnej czy wyszukiwania elementów odstających i anomalii, np. w danych oceny ryzyka.

Po określeniu powodu, a zatem i zakresu łatwiej ustalić rodzaj danych oraz wskaźników oceniających skuteczność modelu. Następnym etapem jest właściwe *zdobycie danych.* W ujęciu technicznym może to oznaczać zgromadzenie danych, zbudowanie potoku zarządzania nowymi danymi, służącego do ciągłego wydobywania danych ze źródła takiego jak strona internetowa czy system CRM, lub po prostu uzyskanie dostępu

do bazy danych istniejącej już w ramach organizacji. Mogą pojawić się tu pewne problemy. Aby wytrenować dobry model, potrzebujemy dobrych danych. Nasze dane natomiast mogą nie spełnić wymogów dotyczących ilości bądź jakości. Mogą także występować **dane osobiste** (ang. *Personally Identifiable Information* — **PII**), które należy maskować albo wykluczyć jeszcze przed rozpoczęciem pracy na danych.

Przy założeniu, że to danetyk pozyskuje dane, może on w końcu użyć swojej wiedzy do zbudowania modelu. Aby to zrobić, potrzebne są dane i algorytm. Uzyskane dane mogą wymagać jakiejś formy przetwarzania. Możemy chcieć wyszukiwać tendencyjność w danych, uzupełniać brakujące wartości lub przekształcać dane tak, żeby były bardziej przydatne dla naszego modelu. Faza ta jest nazywana **przetwarzaniem wstępnym** (ang. *pre-processing*) lub **inżynierią cech** (ang. *feature engineering*). Zadaniem tej fazy jest uzyskanie cech stanowiących dane wejściowe dla naszego modelu.

Po uzyskaniu zestawu cech, często w formie zmiennych uformowanych jako kolumny tabeli, można w końcu *wytrenować model*. Oznacza to, że wypróbujemy algorytmy i ocenimy wyuczone modele na podstawie uzyskanych wskaźników wybranych w oparciu o model. Ta faza jest w swojej naturze iteracyjna i może wymagać wytrenowania wielu modeli (czasami równolegle). Po ocenieniu modelu, na podstawie wymogów określonych w fazie określania powodu korzystania z SI, możemy cofnąć się co najmniej jedną fazę w celu ponownego określenia powodów, zdobycia innych danych lub zmiany decyzji podjętych podczas inżynierii cech.

Gdyż już uznamy, że wytrenowaliśmy wystarczająco dobry model, możemy przejść do ostatniej fazy. To, co się będzie teraz działo, zależy w dużej mierze od sposobu przetwarzania wniosków uzyskanych przez model. Jednym ze sposobów *zintegrowania modelu* jest aplikacja kliencka, gdzie dane są generowane lub gromadzone w tej aplikacji oraz przesyłane do modelu w celu uzyskiwania przewidywań w czasie rzeczywistym, które są również wykorzystywane w aplikacji. Innym powszechnym przykładem jest wykorzystywanie modelu w analizie wsadowej danych. W tym przypadku możemy zintegrować model z potokiem zarządzania danymi, dzięki czemu będziemy w stanie przetwarzać znaczne ilości danych. Bez względu na to, czy dane będą generowane w czasie rzeczywistym, czy wsadowo, jest to ostatni i kluczowy etap, który należy brać pod uwagę podczas realizowania procesu danetycznego zaprezentowanego na rysunku 1.5.

Rysunek 1.5. Pięć etapów procesu danetycznego

Proces danetyczny nie jest liniowy, ale zrozumienie tych pięciu etapów, które prawdopodobnie będziemy wielokrotnie powtarzać, pomoże nam określać, który etap i kiedy należy realizować. Dobry projekt rozpoczyna się od wyraźnie zdefiniowanych powodów

użycia, następnie pozyskujemy dane, przygotowujemy je za pomocą inżynierii cech, trenujemy z ich wykorzystaniem model, który następnie integrujemy z naszymi aplikacjami lub raportami Power BI.

Integracja SI z Power BI jest szczególnie interesująca dla analityka danych. W następnym podrozdziale postaram się wyjaśnić, dlaczego jest to dobrana para.

Dlaczego powinniśmy korzystać ze sztucznej inteligencji w Power BI?

Pytanie o to, dlaczego powinniśmy korzystać z SI w Power BI, jest dwojakiej natury. Po pierwsze możemy się zastanawiać, dlaczego w ogóle powinniśmy korzystać ze sztucznej inteligencji, a po drugie możemy chcieć ustalić, dlaczego powinniśmy korzystać z funkcji SI w Power BI. Aby odpowiedzieć na pierwsze pytanie, musimy poznać możliwości sztucznej inteligencji, czym zajęliśmy się w poprzednich podrozdziałach. Aby poznać odpowiedź na drugie pytanie, musimy zrozumieć, dlaczego SI nie jest jeszcze przyswojona przez większość organizacji.

Problemy z implementacją sztucznej inteligencji

Niewątpliwie istnieje olbrzymie zainteresowanie wszelkimi tematami związanymi ze sztuczną inteligencją. Niestety, podobnie jak w przypadku wszelkich nowych technologii, wszyscy uwielbiają o niej mówić, ale tylko nieliczni rzeczywiście z niej korzystają. Istnieje wiele powodów, dla których przystosowywanie SI przebiega wolniej niż przewidywano (ankieta *McKinsey Survey on AI Adoption* z 2018 r., *https://www.mckinsey.com/featured-insights/artificial-intelligence/ai-adoption-advances-but-foundational-barriers-remain*). Najbardziej oczywistym powodem zdaje się *brak umiejętności*. W poprzednich podrozdziałach wyjaśnialiśmy, czym jest SI i jak możemy tworzyć modele. Wiemy, że w celu wytrenowania modelu musimy wybrać algorytm, a to wymaga wiedzy z zakresu danetyki, która między innymi stanowi połączenie matematyki i statystyki. W konsekwencji ważnym powodem, dla którego firmy nie korzystają z SI, jest fakt, że nie mają pracowników mających umiejętność budowania tych modeli oraz znajomości kryjącego się za tym aparatu matematycznego.

Nie jest to największa przeszkoda. Wielu producentów oprogramowania dostrzegło ten problem już jakiś czas temu i przygotowało narzędzia o usługi przeznaczone dla **danetyków społecznościowych** (ang. *citizen data scientist*), udostępniając technologię szerszemu gronu odbiorców oraz każdemu, kto chciałby z niej korzystać, bez względu na to, czy jest zawodowym danetykiem. Takie przystępne narzędzia nie powinny zastępować inwestycji w SI, ale skłaniają do zadania pytania, dlaczego sztuczna inteligencja nie jest wykorzystywana częściej.

Odpowiedź jest taka, że *ludzie nie znają się na niej*. Ma to związek nie tyle z zatrudnianiem rzeczywistych danetyków, co raczej ze świadomością organizacji. Na wyższych szczeblach kierownictwa nie wiadomo, jak kształtować jasną strategię lub praktyczną wizję wokół SI. Z kolei na niższych szczeblach pracownicy nie potrafią korzystać ze sztucznej inteligencji w codziennej pracy, a nawet jeśli dostarczane są im wnioski uzyskiwane przez SI, często kwestionują oni opinię komputera w przeciwieństwie do własnej intuicji. Wydaje się, że zarówno w przypadku kierownictwa, jak i pracowników problem leży w *niepełnym zrozumieniu możliwości sztucznej inteligencji* oraz w nieznajomości związanych z nią możliwości i ograniczeń.

Nawet kiedy firmy dostrzegają potencjał, przygotowują jasną strategią implementacji SI i zatrudniają w tym celu właściwych ludzi, natrafiają na przeszkody. Gdy pytasz danetyków o najtrudniejszą część ich pracy, rzadko wymieniają *trenowanie modelu*. Największa bolączka dotyczy danych. Z powodów politycznych lub technicznych danetycy *nie mogą pozyskiwać właściwych danych*. Potrzebne dane mogą nawet nie istnieć. A nawet jeśli uzyskają dostęp do właściwych danych, często są one złej jakości lub jest ich zbyt mało, aby wytrenować za ich pomocą dobry model.

Ponadto do faktycznego zaimplementowania sztucznej inteligencji w procesach biznesowych konieczni są nie tylko danetycy, ale cały zespół. Inżynierowie danych są potrzebni do wspomagania wydobywania danych ze źródeł, wielkoskalowego ich oczyszczania oraz przekazywania ich danetykom, którzy mogą dzięki nim trenować modele. Po wytrenowaniu modelu musisz sprawić, żeby spostrzeżenia były przekazywane pionowi biznesowemu w intuicyjny sposób, co stanowi zadanie dla inżynierów oprogramowania integrujących model z aplikacjami klienckimi lub dla analityków danych wizualizujących wnioski płynące z modelu w raportach Power BI.

Innymi słowy w celu utworzenia całościowego rozwiązania i zaimplementowania SI w przedsiębiorstwie potrzebne jest podejście interdyscyplinarne oraz *współpraca między różnymi działami*, w których najlepiej by było, żeby każdy członek miał podstawową wiedzę z zakresu SI, aby budować zaufanie i ułatwiać zmiany.

Dlaczego rozwiązaniem jest sztuczna inteligencja w Power BI?

Skoro rozumiemy już problem, możemy domyślać się, jakie może być jego rozwiązanie. Istnieje wiele powodów powolnego wdrażania SI, dlatego poniżej przedstawiam drobne podsumowanie:

- brak umiejętności danetycznych;
- niepełne zrozumienie koncepcji SI;
- za mało danych lub dane złej jakości;
- brak współpracy między poszczególnymi działami.

Niestety znalezienie utalentowanych danetyków to nie lada wyczyn. Ewentualnie możesz wysłać pracowników na szkolenie, co pomogłoby również w kwestii niepełnego zrozumienia SI w różnych warstwach organizacji. Aby być w stanie gromadzić lepsze dane lub więcej danych, muszą oni wiedzieć, po co to robią. Dlaczego powinniśmy w to inwestować i jakie będą z tego zyski? Z kolei do stymulowania współpracy między działami wymagane są zrozumienie i zaufanie.

Jednym z narzędzi mających wpływ na każdy z tych aspektów jest Power BI. Firmy mają znacznie więcej analityków danych, którzy potrafią pracować z danymi i już korzystają z Power BI lub szybko przestawią się na to narzędzie, niż danetyków. Oznacza to, że analitycy danych znają już wagę dobrych jakościowo danych i mają do nich dostęp. Za pomocą Power BI próbujemy opowiedzieć historię zawartą w spostrzeżeniach generowanych przez dane, dzięki której ludzie mogą podejmować decyzje na podstawie danych. Analitycy danych potrafią przekuwać liczby w intuicyjne fakty. Pomagają przetworzyć wynik uzyskany przez SI w wiarygodne informacje, zrozumiałe dla każdego w organizacji i poza nią. Oznacza to w konsekwencji również ułatwienie współpracy między działami, gdyż Power BI może być już używany w każdym z nich.

Jedyna przeszkoda polega na tym, że osoby korzystające z Power BI często nie są zaznajomione ze sztuczną inteligencją. Mogą one nie mieć wiedzy danetyków, ale mogą bardzo wydajnie pracować na danych. Łącząc SI z Power BI, możemy edukować innych, pomagać tworzyć jasne strategie SI, a także wzbudzać zaufanie użytkowników końcowych do wyników modelu oraz wyjaśniać im, jak to pomaga w procesach biznesowych.

Właśnie dlatego w tej książce są uwzględniane różne opcje SI w Power BI. Obejmują one łatwe rozwiązania, z którymi możesz rozpocząć pracę już dzisiaj i które służą ukazaniu możliwości sztucznej inteligencji. Jednak Power BI można także integrować z zaawansowanymi modelami wytrenowanymi przez danetyków. Dlatego rozsądnym punktem początkowym jest dostosowanie SI w większej skali w ramach Twojej organizacji.

Skoro już wiemy, dlaczego warto korzystać ze sztucznej inteligencji w Power BI, przyjrzyjmy się dostępnym możliwościom.

Jakie mamy opcje sztucznej inteligencji w Power BI?

Jeżeli zastanowimy się, co możemy robić ze sztuczną inteligencją w Power BI, to dostępne opcje możemy podzielić na dwie ogólne kategorie. Po pierwsze dysponujemy łatwymi projektami, z którymi możemy od razu rozpocząć pracę. Po drugie możemy tworzyć własne modele i integrować je z naszymi raportami Power BI, dzięki czemu zyskujemy większą swobodę kosztem nakładu czasu.

Gotowe rozwiązania

Proste rozwiązania sztucznej inteligencji w Power BI można także określać jako **gotowe** (ang. *out-of-the-box*) funkcje SI. Modele te zostały przygotowane przez firmę Microsoft, co oznacza, że nie musimy poświęcać czasu na zbieranie danych potrzebnych do wytrenowania modelu ani mieć wiedzy eksperckiej, aby wybrać właściwy algorytm. Pozwala nam to zaoszczędzić mnóstwo czasu na najbardziej wymagających fazach procesu danetycznego!

W przypadku większości tych funkcji modele są już zintegrowane z Power BI, a nam pozostaje jedynie z nich skorzystać. W niektórych przypadkach mamy możliwość dodania odrobiny własnych danych, aby dostosować model do naszego scenariusza biznesowego. Oznacza to, że dostępny jest jakiś model podstawowy, który został już wytrenowany przez Microsoft na zebranych przezeń danych (sprawdź politykę prywatności (*Privacy Agreement*) dla usługi, w której pracujesz, aby sprawdzić, czy będą wykorzystywane Twoje dane). Dodajemy własne dane, dzięki którym Microsoft może dokończyć uczenie modelu w ułamku czasu, jaki zajęłoby nam stworzenie takiego samego modelu od podstaw.

Ponadto modele te są dostępne na wiele różnych sposobów. Niektóre modele zintegrowane z Power BI są dostępne poprzez bogate wizualizacje, zwane wizualizacjami SI. Istnieją też zintegrowane modele, których można używać wraz ze specyficznymi typami danych, na przykład możemy korzystać z prognoz podczas przetwarzania szeregów czasowych. Możemy także korzystać z zestawu usług **Cognitive Services** na platformie Azure, składającego się z gotowych modeli, które można łatwo zintegrować z dowolną aplikacją za pomocą interfejsów API.

Tworzenie własnych modeli

Zaletą korzystania z gotowych modeli jest oszczędność czasu i pieniędzy podczas pierwszych kroków z SI. Wadą używania takich modeli udostępnionych przez producentów oprogramowania jest mniejsza kontrola i swoboda przy projektowaniu modelu. Jeżeli wolisz tworzyć własne modele, potrzebujesz dostępu do wiedzy danetycznej, aby podejmować właściwe decyzje podczas trenowania modelu.

Mimo to wciąż występuje wiele sytuacji, w których chcemy mieć pewność, że projektowany model będzie skrojony pod nasze potrzeby. W tej książce założymy, że chcemy pracować z platformą usług rozproszonych Azure firmy Microsoft w celu łatwej integracji dowolnego modelu uczenia maszynowego z Power BI. Mamy trzy główne możliwości w platformie Azure podczas trenowania modelu:

- korzystanie ze zautomatyzowanego uczenia maszynowego (**Automated Machine Learning**) do trenowania (równoległego) wielu modeli, wyboru najlepszego z nich i jego integracji z naszym potokiem danych;

- korzystanie z projektanta uczenia maszynowego Azure (**Azure Machine Learning Designer**) do stworzenia modelu.

  Obydwa te rozwiązania wymagają mniejszej wiedzy danetycznej niż w przypadku trzeciej możliwości:

- korzystanie z przestrzeni roboczej uczenia maszynowego Azure (**Azure Machine Learning workspace**) do trenowania i wdrażania modelu na podstawie napisanych od podstaw skryptów uczących; są one napisane w językach Python lub R i najczęściej trenują modele za pomocą takich otwartych bibliotek, jak **Scikit-Learn**, **PyTorch** i **TensorFlow**.

Bez względu na wybraną opcję celem jest uzyskanie pełnej kontroli nad danymi uczącymi oraz możliwość wyboru algorytmów stosowanych podczas uczenia modelu. Oznacza to, że może być potrzebne więcej umiejętności, czasu i mocy obliczeniowej do uzyskania takich samych rezultatów jak w przypadku gotowych modeli. Zatem zarówno modele gotowe, jak i budowane od podstaw spełniają swoje zadanie i w każdym przypadku stosowania należy ocenić, które rozwiązanie sprawdza się najlepiej.

Podsumowanie

W tym rozdziale zastanawialiśmy się, jakie umiejętności cechują dobrego analityka danych, poznaliśmy podstawowe informacje na temat SI, dowiedzieliśmy się, dlaczego połączenie SI z Power BI pomaga w przystosowywaniu SI, a także przyjrzeliśmy się zagadnieniom opisywanym w dalszej części książki, mianowicie dostępnym opcjom SI w Power BI. Nie wszystkie z tych opcji będą istotne dla Ciebie lub Twojej organizacji, ale kolejne rozdziały przynajmniej dadzą Ci pogląd na to, co jest możliwe, dzięki czemu będziesz mógł uczyć innych i siebie. Mam nadzieję, że w ten sposób będziesz pojmować sztuczną inteligencję jako pomocną dla Twojej firmy w stworzeniu jasnej strategii, która przekształci organizację tak, by była bardziej nastawiona na dane i wykorzystywała SI na dużą skalę.

W następnym rozdziale skoncentrujemy się na pierwszym elemencie, którego potrzebujemy do uczenia modelu: danych. Nauczymy się eksplorować dane, dzięki czemu zrozumiemy, jak powinny wyglądać dane wejściowe oraz co należy zrobić, aby zestaw danych nadawał się do naszego zadania.

Przekonaj się, jaki potencjał tkwi w analizie zbiorów danych!

Microsoft Power BI zdobył uznanie jako znakomite narzędzie do analizy i przetwarzania złożonych zbiorów danych, ale to nie koniec jego możliwości. Power BI nadaje się do wydobywania z modeli sztucznej inteligencji informacji, które mogą się stać wartościowym materiałem wspomagającym podejmowanie najlepszych decyzji biznesowych. Aby jednak w pełni skorzystać z funkcji dostępnych w Power BI, trzeba posiąść podstawową wiedzę o sztucznej inteligencji.

Książka stanowi wprowadzenie do pracy z funkcjami SI dostępnymi w Power BI; jest skierowana do osób znających to środowisko. Dowiesz się z niej, w jaki sposób sztuczna inteligencja może być używana w Power BI i jakie funkcje są w nim domyślnie dostępne. Nauczysz się też eksplorować i przygotowywać dane do projektów SI. Pokazano tu, jak umieszczać dane z analizy tekstu i widzenia komputerowego w raportach Power BI, co ułatwia korzystanie z zewnętrznej bazy wiedzy. Omówiono również procesy tworzenia i wdrażania modeli AutoML wytrenowanych na platformie Azure ML, a także umieszczania ich w edytorze Power Query. Nie zabrakło kwestii związanych z prywatnością, bezstronnością i odpowiedzialnością w korzystaniu z SI.

W książce między innymi:

- unikanie tendencyjności w przetwarzaniu danych
- szeregi czasowe i prognozowanie w Power BI
- wykrywanie anomalii
- analiza tekstu w Power Query
- trenowanie własnych modeli
- integracja Azure ML z Power BI i generowanie przewidywań

Mary-Jo Diepeveen pracuje w Microsofcie. Kiedyś interesowała się neurobiologią, szczególnie nieświadomymi wzorcami zachowań, i ta wiedza okazała się pomocna w zrozumieniu niektórych aspektów uczenia maszynowego. Obecnie koncentruje się na tworzeniu treści edukacyjnych w zakresie danetyki i sztucznej inteligencji.

Helion

helion.pl

HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

KOD KORZYŚCI
Sięgnij po więcej!

ISBN 978-83-8322-780-1

Cena: 79,00 zł

Packt>