O książce

Model uczenia maszynowego z dużą dokładnością przewiduje postępy leczenia pacjentów – to jeden z wielu podobnych nagłówków prasy popularno-naukowej, znaleziony w czasopiśmie *Nature* podczas pisania tego wstępu [8]. Książka uczy, jak samodzielnie budować takie światowej klasy modele uczenia maszynowego i jak wdrażać gotowe modele do użycia.

Modele uczenia maszynowego mogą być używane do predykcji, opisywania wzorców ukrytych w danych (deskrypcji) oraz do kompresji i generowania danych (np. do generowania obrazów lub opisów sekwencji wideo). Książka koncentruje się na najpopularniejszych modelach predykcyjnych, czyli modelach stosujących wykryte w treningowych danych wzorce do uzupełniania brakujących danych o nowych przypadkach, niewidzianych przez model podczas treningu.

Dla kogo jest ta książka?

Rynek sztucznej inteligencji rośnie na tyle szybko, że specjalistów od przetwarzania danych ciągle brakuje. Inżynieria danych (ang. *data science*) to interdyscyplinarna wiedza, której opanowanie wymaga znajomości algebry, geometrii, statystyki, rachunku prawdopodobieństwa i algorytmiki, uzupełniona o praktyczną umiejętność programowania w przynajmniej dwóch z trzech najpopularniejszych językach danych: SQL, R lub Python. Co więcej, sztuczna inteligencja jest przedmiotem intensywnych badań naukowych i samo śledzenie postępów w tej dziedzinie wiąże się z regularnym (codziennym) dokształcaniem. Nic dziwnego, że inżynierowie danych są jednymi z najbardziej pożądanych i najlepiej wynagradzanych pracowników.

Zbudowanie modelu uczenia maszynowego wymaga:

- specjalistycznej wiedzy z dziedziny, w ramach której projekt jest realizowany (np. medycyny czy logistyki transportu); prawie zawsze wymaga to wsparcia eksperta z danej dziedziny;

- praktycznej znajomości statystki i umiejętności wizualizacji danych niezbędnej do oceny jakości danych;
- praktycznej znajomości języka SQL, R lub Python niezbędnej do uporządkowania, wstępnego przygotowania i wzbogacenia danych;
- zrozumienia zasad działania poszczególnych algorytmów uczenia maszynowego koniecznych do ich wyboru i optymalizacji (do czego przyda się z kolei znajomość algebry i geometrii);
- użycia języka R lub Python do stworzenia, oceny, zoptymalizowania i wdrożenia do produkcji modeli eksploracji danych (do oceny jakości modeli ponownie przyda się znajomość statystyki uzupełniona o wiedzę z zakresu rachunku prawdopodobieństwa).

Książka jest adresowana do wszystkich, którzy chcieliby poznać lub udoskonalić swoją wiedzę z powyższych (z wyjątkiem pierwszego) obszarów. Tego typu książki mogą być przystępne, napisane językiem potocznym i ilustrowane praktycznymi przykładami lub dokładne – pełne precyzyjnych równań matematycznych. Z założenia książka ma być przystępna, co oznacza, że opisowe wyjaśnienia pozostawiają Czytelnikowi możliwość ich różnorodnego interpretowania. Problem ten starałem się rozwiązać, ilustrując opisywane zagadnienia praktycznymi przykładami, których samodzielne wykonanie powinno rozwiać ewentualne niejasności. Takie podejście ma tę dodatkową zaletę, że kładzie nacisk na cenniejsze od wiedzy teoretycznej umiejętności praktyczne [9].

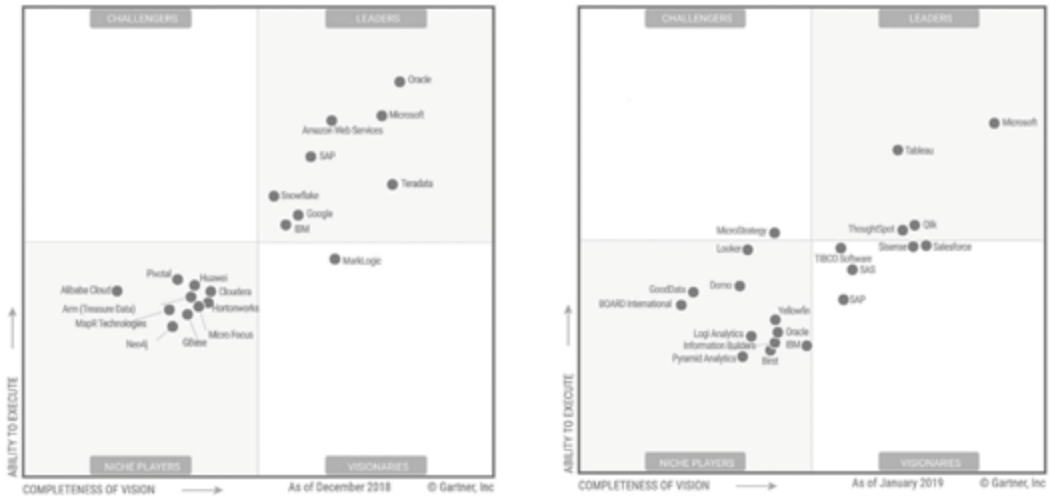
Liczę, że dzięki temu trafię do szerokiego grona Czytelników i zarówno studenci kierunków informatycznych, jak również analitycy, programiści, administratorzy baz danych oraz statystycy znajdą w książce informacje, które pozwolą im opanować praktyczne umiejętności potrzebne do samodzielnego tworzenia systemów uczenia maszynowego. Na koniec dodam, że umiejętność czytania ze zrozumieniem kodu SQL, R i Python na pewno ułatwi lekturę książki.

Narzędzia

Zilustrowanie opisanych w książce zagadnień praktycznymi przykładami wymagało wyboru jakichś narzędzi. **Wybór padł na serwer SQL Server 2019 i program Power BI Desktop**, ponieważ:

- oba te narzędzia są dostępne za darmo; edycja SQL Server Developer Edition ma pełne wsparcie dla opisanych w książce usług uczenia maszynowego i może być bezpłatnie używana do nauki oraz testów, jedynie produkcyjne wykorzystanie utworzonych modeli będzie wymagało zakupu odpowiedniej licencji; natomiast program Power BI Desktop jest całkowicie darmowy i może być wykorzystywany do dowolnych celów, również komercyjnych;
- SQL Server może być zainstalowany w środowiskach Windows, Linux lub na platformie Docker;
- SQL Server pozwala wydajnie przechowywać duże zbiory danych i przetwarzać je za pomocą języka SQL;

- usługi uczenia maszynowego serwera SQL Server pozwalają tworzyć modele uczenia maszynowego przy użyciu języków R, Python lub Java;
- integracja języków R, Python i Java z serwerem SQL Server wykracza poza prostą możliwość uruchamiania skryptów tych języków po stronie serwera i pozwala nie tylko wydajnie analizować dane, lecz także wdrażać gotowe modele uczenia maszynowego do użycia;
- usługa Power BI i program Power BI Desktop są kompletnymi narzędziami do tworzenia samoobsługowych systemów BI; pozwalają one w prosty sposób pobrać dane z różnych źródeł, dowolnie je przekształcać, tworzyć rozbudowane modele biznesowe i interaktywne, rozbudowane wizualizacje;
- agencja Gartnera od wielu lat wysoko pozycjonuje oba te produkty w swoich corocznych raportach (rys. 1).



Rysunek 1. Po lewej stronie – raport dotyczący serwerów baz danych w zastosowaniach analitycznych (Gartner's 2018 Magic Quadrant for Data Management Solutions for Analytics), po prawej – raport przedstawiający platformy BI (Gartner's 2019 Magic Quadrant for Analytics and Business Intelligence Platforms)

Przykłady

W tekście książki zostały opisane tylko wybrane fragmenty kodu użytego do ilustracji omawianych zagadnień. Kompletną wersję przykładów razem z użytymi danymi można pobrać ze strony Wydawnictwa, a ich ostatnią wersję z serwisu Github. Całość repozytorium możecie Państwo pozyskać wieloma sposobami, jednym z nich jest zastosowanie narzędzia Git, dostępnego w systemach Windows oraz Linux. Przykładowo, korzystając z polecenia:

```
$ git clone https://github.com/szelor/practical-machine-learning.
git
```

Pod adresem `https://it.pwn.pl/Artykuly/Praktyczne-uczenie-maszynowe-materialy-dodatkowe` znajdą Państwo archiwum ZIP z kopią bazy danych serwera SQL Server 2019 zawierającą oprócz tabel z danymi, widoki, procedury składowane i funkcje potrzebne do utworzenia opisywanych modeli uczenia maszynowego. Archiwum to zawiera również podzielone między foldery pliki z danymi, pliki Power BI Desktop i skrypty w językach R i Python, które pomogą Państwu wykonać opisywane projekty.

Pod adresem `https://github.com/szelor/practical-machine-learning` znajdą Państwo podzielone między foldery repozytorium zawierające pliki z danymi, pliki Power BI Desktop i skrypty w językach R i Python, które pomogą Wam wykonać opisywane projekty. W repozytorium nie znajdują jednak Państwo kopii przykładowej bazy danych, zbyt dużej, żeby można było ją tu udostępnić.

Najprościej jest skorzystać z gotowych plików i w ten sposób przekonać się, jak działają poszczególne modele, skorzystać z interaktywnych wizualizacji danych i spojrzeć na kolorowe wykresy w wysokiej rozdzielczości. Zachęcam jednak do wspólnego rozwoju tego projektu – każdy może modyfikować i rozbudowywać te skrypty, a następnie dzielić się wynikami swojej pracy z innymi.

Bibliografia

Do książki została dołączona bibliografia. Starałem się umieścić w niej jak najwięcej odnośników do ogólnodostępnych, elektronicznych wersji wymienionych w niej pozycji. Oczywiście, jeżeli jakaś pozycja nie jest dostępna za darmo, odnośnika do niej nie ma. Namawiam do zapoznania się z tymi artykułami i książkami – zarówno tymi klasycznymi, pochodzącymi z lat 70. XX w., jak i opisującymi wyniki najnowszych badań w obszarze uczenia maszynowego.

Konwencje i oznaczenia

W książce zostały zastosowane następujące konwencje do oznaczania różnych typów tekstu:

- **czcionką pogrubioną są wyróżnione** nowe, istotne zagadnienia, na które czytelnik powinien zwrócić uwagę;
- *czcionką pochyłą* są pisane nazwy własne w miejscach ich pierwszego użycia oraz angielskie odpowiedniki wprowadzanych pojęć;
- `czcionką stałej szerokości` znaków są pisane przykładowe programy, pojawiające się w treści akapitu fragmenty programów (instrukcje, słowa kluczowe, modyfikatory itp.), polecenia wprowadzane z klawiatury, teksty wyświetlane na ekranie oraz adresy internetowe;
- uwagi, wskazówki, ciekawostki, dobre rady lub ostrzeżenia są pisane

►► mniejszym stopniem pisma i wyróżnione znakiem widocznym obok.