

PODSTAWY UCZENIA MASZYNOWEGO DLA MENADŻERÓW PROJEKTU IT

Patryk MŁODZIANOWSKI^{1*}, and Robert ROSTKOWSKI²

¹ Państwowa Uczelnia Zawodowa im. Ignacego Mościckiego w Ciechanowie;
patryk.mlodzianowski@puzim.edu.pl; ORCID: 0000-0002-2404-4329

¹ Państwowa Uczelnia Zawodowa im. Ignacego Mościckiego w Ciechanowie, Koło Naukowe Informatyki
Stosowanej; robertrostkowski@o2.pl; ORCID: 0000-0003-2146-8131

* Korespondencja: patryk.mlodzianowski@puzim.edu.pl

Streszczenie: Celem artykułu jest agregacja i podsumowanie wiedzy specjalistycznej z zakresu uczenia maszynowego oraz przedstawienie jej w sposób przystępny dla osób zarządzających projektami, w których wykorzystywane jest uczenie maszynowe. Specyfika tego obszaru wymaga znajomości podstawowych pojęć, aby sprawnie komunikować się z zespołem. Zrozumienie problematyki uczenia maszynowego jest kluczowe z punktu widzenia całego projektu, ponieważ pozwala lepiej i bardziej precyzyjnie zarządzać ryzykiem w projekcie. W tekście dokonano analizy literatury specjalistycznej zarówno z branży IT jak i zarządzania projektami. Następnie dokonano jej syntezy. Artykuł przedstawia wybrane, interdyscyplinarne aspekty wytwarzania oprogramowania wykorzystującego sztuczną inteligencję. Podkreślono także istotne różnice pomiędzy wytwarzaniem oprogramowania klasycznego a oprogramowania opartego o uczenie maszynowe. Dokonano także analizy potencjalnego ryzyka projektowego jakie może wiązać się z obszarem sztucznej inteligencji.

Słowa kluczowe: sztuczna inteligencja, uczenie maszynowe, zarządzanie projektami, zarządzanie ryzykiem, szacowanie ryzyka

THE BASICS OF MACHINE LEARNING FOR IT PROJECT MANAGERS

Abstract: The main purpose of this article is to summarize specialist knowledge in the field of machine learning and present it in an accessible manner for project managers who work on a project that use machine learning. The specificity of this area requires knowledge of basic concepts in order to communicate effectively with the team. Understanding the issues of

machine learning is crucial for the entire project, because it allows to prepare better risk management models. The text analyzes specialist literature from both the project management and IT industry. The article presents selected interdisciplinary aspects of software development using artificial intelligence. Significant differences between the production of classical software and machine learning based software were also emphasized. The potential risk that may be associated with the area of artificial intelligence was also analyzed

Keywords: artificial intelligence, machine learning, project management, risk management, risk assessment

1. Wstęp

Od wielu lat branża IT nieustannie się rozwija, dzięki czemu obserwować można dynamiczny postęp technologiczny. Ma to bezpośrednie przełożenie na niemalże wszystkie gałęzie gospodarki, które dzięki szeroko rozumianej informatyzacji mogą upraszczać, ulepszać czy automatyzować swoje procesy (przemysł 4.0). Owa informatyzacja dostarcza nowych danych i informacji, które pozwalają budować nowsze rozwiązania informatyczne, aby jeszcze lepiej optymalizować procesy przedsiębiorstw. Jest to swego rodzaju cykl, który z każdym obiegiem wyczerpuje możliwości klasycznego oprogramowania i wymaga wykorzystania bardziej wyrafinowanych technologii.

Ogromny przyrost danych w ostatnich dekadach spowodował, że na rynku IT pojawił się nowy sektor – Big Data, zajmujący się analizą dużych zbiorów danych. Zaraz obok niego pojawił się także bliźniaczy sektor dotyczący uczenia maszynowego, który zajmuje się wykorzystaniem tych danych do opracowywania modeli wykorzystywanych w oprogramowaniu. Obecnie zaobserwować można intensywny wzrost ilości oprogramowania, bazującego na sztucznej inteligencji. Można także z dużym prawdopodobieństwem stwierdzić, że większość wytwarzanych programów wykorzystuje w mniejszym lub większym stopniu rozwiązania związane z tym obszarem. Efektem tego jest stale zwiększające się zapotrzebowanie nie tylko na personel techniczny ale także kierowniczy, który będzie znał i rozumiał specyfikę oraz problematykę wykorzystania nowoczesnych metod sztucznej inteligencji, takich jak uczenie maszynowe.

Należy zaznaczyć, że sztuczna inteligencja (ang. artificial intelligence) jest pojęciem dość starym i bardzo ogólnym, które zawiera także metody i algorytmy oprogramowania klasycznego.

2. Zastosowania uczenia maszynowego

Jak podkreślono wcześniej - machine learning i sztuczna inteligencja nie są dokładnie tym samym. Uczenie maszynowe jest podzbiorem sztucznej inteligencji. Rolą tego podzbioru jest uczenie maszyn pewnych reguł postępowania i stosowania ich do rozwiązywania danego problemu, zamiast bezpośredniego wskazywania sposobu rozwiązania. Głębokie uczenie jest podzbiorem uczenia maszynowego, które charakteryzuje się wykorzystaniem sztucznych sieci neuronowych.



Rysunek 4. Schemat obszarów sztucznej inteligencji (źródło: opracowanie własne).

Machine learning jest wykorzystywany w wielu aspektach, głównie do przetwarzania obrazu (np. klasyfikacja obrazów, wykrywanie obiektów), dźwięku (np. konwersja mowy na tekst), tekstu czy predykcji zjawisk (np. przewidywanie opadów, cen produktów, cen akcji na giełdzie). Punktem wejściowym do zagadnień uczenia maszynowego są dane. To one stanowią podstawę, która pozwala, przy zastosowaniu odpowiednich algorytmów, przekazać maszynie informację o charakterystycznych wzorcach, a dzięki temu „nauczyć” ją jak te wzorce wykrywać.

Większość osób w swoim życiu codziennym ma styczność z oprogramowaniem korzystającym z machine learningu, często nie będąc tego świadomymi. Nikogo nie dziwi, że smartfony potrafią rozpoznać twarz właściciela, odtwarzacze filmów mogą wyświetlać napisy na podstawie dźwięku, a na wyciągnięcie ręki dostępne są wielojęzyczne, automatyczne translatory. Wszystkie te przypadki łączy jedno - uczenie maszynowe.

Bardzo mocną domeną tego obszaru jest także prognozowanie. Z powodzeniem można dość precyzyjnie przewidywać różne zjawiska. Koronnym przykładem są modele analizujące

dane finansowe (np. do prognozowania kursu walut), giełdowe (np. do prognozowania cen akcji) czy marketingowe (np. do rekomendowania produktów klientom).

3. Uczenie maszynowe – ogólna charakterystyka

Uczenie maszynowe jest częścią sztucznej inteligencji, której zadaniem jest przygotowanie modelu pozwalającego na podejmowanie decyzji lub przewidywanie zjawisk w oparciu o analizę danych dostarczonych na etapie uczenia. Dane te nazywane są danymi treningowymi (ang. training data) i stanowią fundament w tworzeniu modelu. Oczywiście równie istotny jest dobór algorytmu, który przeprowadzi uczenie. Wybór optymalnego algorytmu wymaga wcześniejszego sprecyzowania planowanego zastosowania modelu oraz dokonania analizy potencjalnych danych treningowych. Niezwykle ważna jest ich jakość oraz ilość.

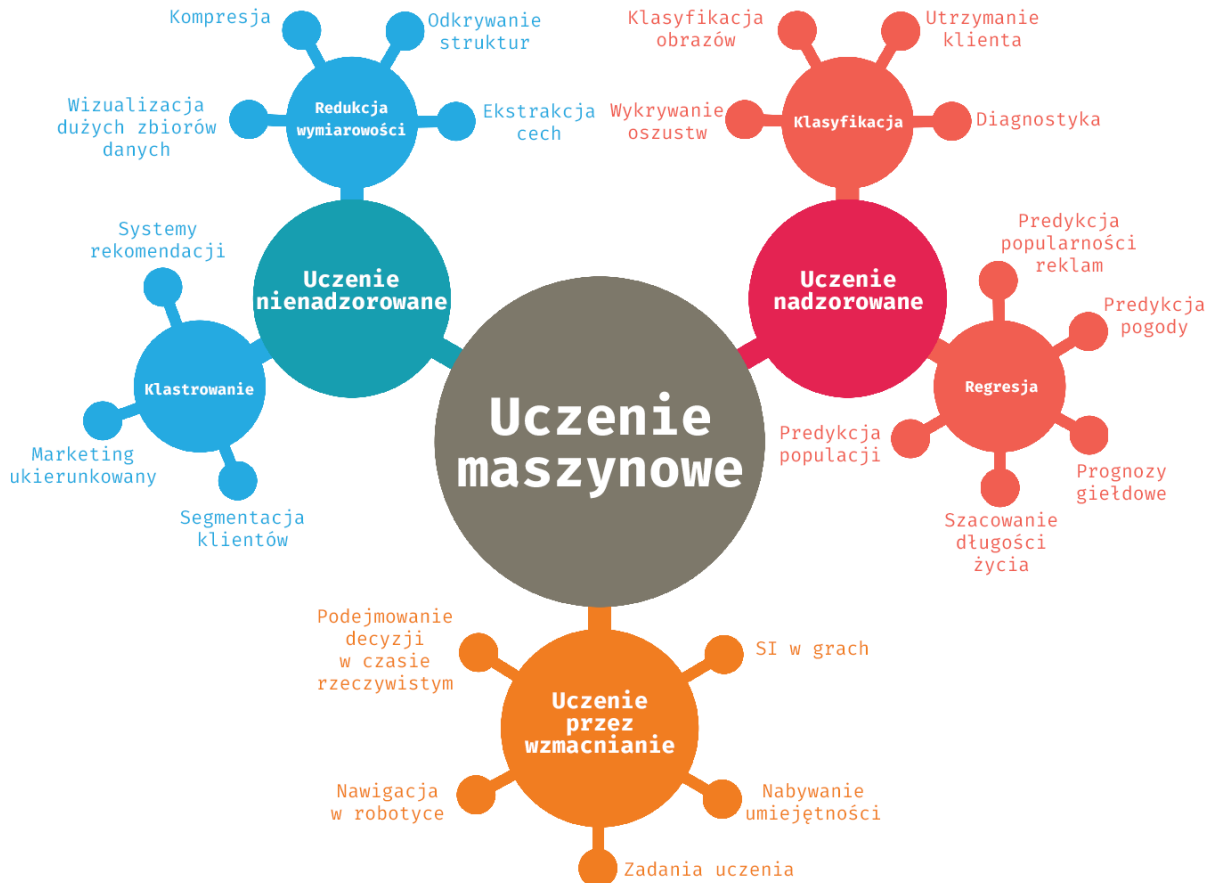
Wyróżnić można trzy rodzaje uczenia maszynowego względem kryterium sposobu uczenia.

- Nadzorowane (ang. supervised learning) – wykorzystuje relację pomiędzy danymi wejściowymi i oczekiwaną wartością. Wymaga dostarczenia zbioru danych wejściowych wraz z informacją o spodziewanym wyniku. W przypadku klasyfikacji obrazów pojedynczą daną może być zdjęcie wraz z informacją o tym co się na nim znajduje. W procesie uczenia algorytm spróbuje wydobyć cechy z dostarczonych danych i ustalić ich wagi.
- Nienadzorowane (ang. unsupervised learning) – ten rodzaj uczenia maszynowego nie wymaga podawania oczekiwanego wyniku. Zadaniem algorytmów jest odkrycie wzorców i pogrupowanie danych lub redukcja liczby wymiarów.
- Uczenie przez wzmacnianie (ang. reinforcement learning) – polega na działaniu agenta w środowisku, z którego będzie czerpał dane na temat dobrych lub złych wyborów. Głównym celem jest maksymalizacja nagrody zdobywanej przez agenta. W ten sposób model uczy się podejmować lepsze decyzje na bazie analizy wyników decyzji podjętych wcześniej.

Ze względu na sposób rozwiązywania problemów przez machine learning, wyróżnić można klasyfikację, regresję, klastrowanie, czy redukcję wymiarowości.

- Klasyfikacja – polega na utworzeniu modelu, który zaklasyfikuje dane wartości do jednej z klas znanych z procesu uczenia. Filtry spam/nie spam są przykładem klasyfikacji binarnej, czyli takiej, w której dostępne są tylko dwie klasy. Wieloklasowa klasyfikacja pozwala na tworzenie modeli uwzględniających więcej niż dwie klasy.
- Regresja – modele regresji korzystają z metod statystycznych. Ich zadaniem jest znalezienie funkcji, która będzie najlepiej pasowała do danych uczących.

- Klastrowanie – polega na grupowaniu danych względem zbliżonych atrybutów. Jest często wykorzystywane w uczeniu nienadzorowanym.
- Redukcja wymiarowości – jest także wykorzystywana w uczeniu nienadzorowanym. Główną ideą jest uproszczenie zbioru danych i redukcja kompleksowości.



Rysunek 5. Podzbiory uczenia maszynowego (źródło: <https://www.freecodecamp.org/news/machine-learning-for-managers-what-you-need-to-know>).

Największą wadą modeli wytworzonych przez uczenie maszynowe jest to, że nigdy nie będą doskonałe. Efektywność modeli może być bliska 100%, ale nigdy nie powinna być równa 100%. Wytrenowane modele zwykle określają prawdopodobieństwo wystąpienia danej wartości na podstawie danych wejściowych. Ta wartość, która jest najbardziej prawdopodobna zostaje podana jako wynik. Przy takim sposobie działania, oczywistym jest, że model niekiedy może być zawodny. Jest to nieco inne podejście niż w oprogramowaniu klasycznym, gdzie wynik zwykle jest jednoznaczny i pewny.

4. Kiedy warto wykorzystać uczenie maszynowe w projekcie?

Przed podjęciem ostatecznej decyzji o wdrożeniu rozwiązań typu machine learning do projektu, należy przeanalizować czy jest to uzasadnione oraz czy ryzyko i potencjalne korzyści będą opłacalne.

Istotne jest zrozumienie obecnego procesu, który ma zostać usprawniony w wyniku projektu. Czy projekt ma całkowicie zastąpić pewien proces czy tylko wspomóc?

Niekiedy, w zależności od planowanego zastosowania, możliwe jest znalezienie alternatywnych rozwiązań w statystyce lub innych naukach związanych z analizą danych. Rozsądny dobór rozwiązania adekwatnego do postawionego problemu z pewnością przełoży się na optymalizację projektu.

Ogólne przesłanki biznesowe, przemawiające za tym, że wybór ścieżki uczenia maszynowego jest poprawny są następujące:

- dostępność dużego i usystematyzowanego zbioru danych, na podstawie którego dotychczas podejmowano decyzje,
- zadania bazowały na wytworzonym wcześniej procesie heurystycznym lub jasno zdefiniowanych regułach wynikających z analizy danych.

Przesłanki, które mogą znacznie zwiększyć szansę niespełnienia założeń projektu:

- zadania, które mają zostać zautomatyzowane poprzez wdrożenie uczenia maszynowego, opierają się głównie na cechach ludzkich, które ciężko jest odwzorować w postaci logicznej (np. empatia, kreatywność),
- niewielki lub niereprezentatywny zbiór danych (uczenie powinno odbywać się za pomocą danych, które będą wiernie odzwierciedlały dane, z którymi docelowo pracować ma model).

5. Narzędzia wspomagające prowadzenie projektu

Powszechnie panuje przekonanie, że projekt wykorzystujący uczenie maszynowe może być zarządzany jak klasyczny projekt softwareowy. Niestety różnice w procesie wytwarzania tak specyficznego oprogramowania wymuszają zmianę w samym sposobie zarządzania, a także w narzędziach, które to zarządzanie wspierają.

Wybór narzędzi do prowadzenia projektu powinien być przemyślany. Najlepiej wybrać rozwiązania kompleksowe, dedykowane tego typu projektom. Niestety jest ich o wiele mniej, a ich cena jest często znacznie wyższa. Wyzwaniem może okazać się także dobór odpowiedniej metody prowadzenia projektu. Wynika to z faktu, iż wprowadzenie machine learningu będzie znacząco zmieniało projekt w stosunku do wytwarzania kanonicznego oprogramowania. Jest to spowodowane różnicami w sposobie definiowania pewnych cech projektu, takich jak zarządzanie ryzykiem, śledzenie postępów, szacowanie czasu zadań, czy nawet samo definiowanie zadań. Przez to niektóre metodologie zarządzania projektami

mogą okazać się bardzo kłopotliwe w przypadku takich projektów - np. popularny scrum. Ponadto znacząco różnić się będą wymagania dotyczące archiwizacji kodu źródłowego, rejestru modeli oraz samych zbiorów danych.

Rynek oprogramowania wspierającego prowadzenie projektów IT z naciskiem na zastosowanie uczenia maszynowego, stale się powiększa. Najbardziej znanymi narzędziami są Neptune.ai, MLflow, Sagemaker czy Azure Machine Learning. Pozwalają w przystępny sposób wersjonować dane, prowadzić rejestr modeli, a także śledzić rezultaty eksperymentów. W większości ułatwiają także samą implementację rozwiązań poprzez dostarczenie zintegrowanych środowisk programistycznych, wsparcia w wyborze odpowiednich algorytmów, a także upraszczają proces przygotowania danych i proces uczenia. Dodatkowo umożliwiają wdrożenie gotowego modelu. Większość z nich umożliwia także wykorzystanie mocy obliczeniowej chmury, jednakże należy mieć na uwadze, że zwykle jest to dość kosztowne ze względu na duże zapotrzebowanie uczenia maszynowego na moc obliczeniową.

6. Ryzyko wynikające z wykorzystania uczenia maszynowego w projekcie

Wydawać by się mogło, że po określeniu celu projektu i wybraniu aspektu jaki ma realizować uczenie maszynowe (np. wykrywanie obiektów na obrazie czy predykcja wartości) pozostaje tylko jego realizacja. Jednak przy wszystkich korzyściach jakie daje zastosowanie uczenia maszynowego, warto zastanowić się nad jego wadami. Niestety zbudowanie dobrego modelu wiąże się ze zwiększonym ryzykiem niepowodzenia całego projektu. Zagrożenia, o których należy pamiętać podczas szacowania ryzyka projektowego przedstawiono poniżej w postaci listy.

- Przede wszystkim konieczna jest wykwalifikowana kadra techniczna – programiści, analitycy danych. Niestety w Polsce notowany jest deficyt specjalistów z tej dziedziny.
- Specjaliści z dziedziny, której dotyczy rozwiązywany problem (np. ekspert z zakresu ekonomii w przypadku budowania modelu predykcji cen) – pomagają zespołowi technicznemu zrozumieć istotę danych oraz relacje pomiędzy nimi.
- Warto zastanowić się nad ilością i jakością danych, na bazie których ma powstać model (baza wiedzy dla uczenia). Wiele algorytmów uczenia maszynowego wymaga olbrzymiej ilości danych do treningu.
- Dobór algorytmu uczącego lub opracowanie własnego. Obecnie dostępnych jest wiele algorytmów realizujących uczenie, które można wykorzystać. Jednak ich mnogość może przysporzyć problemów w wybraniu jednego – najlepiej pasującego do zagadnienia, które jest celem projektu.

- Oprócz samych danych, istotna jest jeszcze kwestia tzw. hiperparametrów, które są charakterystyczne dla wybranego algorytmu. Zazwyczaj trudno jest jednoznacznie stwierdzić jakie będą najlepsze ustawienia hiperparametrów dla danego uczenia i zwykle dobiera się je na zasadzie eksperymentowania (trenuje się wiele modeli w różnych wariacjach hiperparametrów i porównuje ich sprawność). Eksperymenty mogą przez długi czas nie przynosić akceptowalnych rezultatów, dlatego trudno jest oszacować czas trwania tej fazy projektu.
- Ogromną rolę odgrywa zaplecze techniczne, ponieważ uczenie maszynowe wymaga dużej mocy obliczeniowej. Może okazać się, że przez zbyt słabe zasoby obliczeniowe nie można wytrenować dobrego modelu w rozsądnym czasie lub znacznie utrudnia to eksperymentowanie z hiperparametrami.
- Nie ma gwarancji, że wytrenowany model jest wystarczająco dobry, aby mógł zostać wdrożony jako rozwiązanie produkcyjne.
- Ponadto należy mieć świadomość, że uczenie maszynowe często nazywane jest czarnym pudłem (ang. black box), ponieważ gotowy model przyjmuje dane wejściowe i oddaje dane wyjściowe, jednak prześledzenie tego co dzieje się pomiędzy tymi etapami jest bardzo trudne, a czasami nawet niemożliwe.

7. Podsumowanie

Prowadzenie projektu, w którym wykorzystywane są techniki uczenia maszynowego stanowi unikalne wyzwanie, nie tylko implementacyjne, ale także zarządcze. Wiąże się to także z koniecznością uwzględniania listy aspektów ryzyka, które są specyficzne dla tego obszaru. Ponadto wymagana jest specjalistyczna wiedza domenowa, które także pomoże wskazać słabe punkty założeń projektu.

Dla tego typu projektu istotne są także odpowiednio dobrane narzędzia zapewniające kompleksową obsługę i usprawniające nadzór nad płynnością postępu prac. Podstawowy wachlarz narzędzi do zarządzania projektem musi bowiem zostać wzbogacony rozwiązaniami dedykowanymi dla uczenia maszynowego.

Dokładne oszacowanie ryzyka pozwala na wczesnym etapie wykryć punkty blokujące czy wręcz uniemożliwiające osiągnięcie założonego celu.

Powodzenie projektu korzystającego z uczenia maszynowego zależy od wielu czynników. Najbardziej istotny jest dobór danych i ich odpowiednie przygotowanie. Gruntowna analiza danych powinna być pierwszym krokiem decydującym o tym czy projekt ma szansę zostać zakończony sukcesem.

Bibliografia

1. Aven, T., and Renn, O. (2010). *Risk Management and Governance*. Springer.
2. Badiru, A.B. (2019). *Project Management*.
3. Dandage, R.V., Mantha, S.S., Rane, S.B., & Bhoola, V. (2018). *Analysis of interactions among barriers in project risk management*. Journal of Industrial Engineering International, 14, 153-169.
4. Duncan, W.R. (1996). *A guide to the project management body of knowledge*.
5. Gharibi, G., Walunj, V., Alanazi, R., Rella, S., & Lee, Y. (2019). *Automated Management of Deep Learning Experiments*. Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning.
6. Hopkin, P. (2018). *Fundamentals of Risk Management: Understanding, Evaluating and Implementing Effective Risk Management*. Kogan Page Publishers.
7. Muriana, C., & Vizzini, G. (2017). *Project risk management: A deterministic quantitative technique for assessment and mitigation*. International Journal of Project Management, 35, 320-340.
8. Park, M. (2020). *A Monitoring System for Machine Learning Models in a Large-Scale Context*.
9. Poth, A., Meyer, B., Schlicht, P., & Riel, A. (2020). *Quality Assurance for Machine Learning – an approach to function and system safeguarding*. 2020 IEEE 20th International Conference on Software Quality, Reliability and Security (QRS), 22-29.
10. Schelter, S., Biessmann, F., Januschowski T., Salinas D., Seufert S., and Szarvas G. (2018). *On Challenges in Machine Learning Model Management*. IEEE Data Engineering Bulletin.
11. Schelter, S., Rukat, T., & Biessmann, F. (2020). *Learning to Validate the Predictions of Black Box Classifiers on Unseen Data*. Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data.
12. Schwarz, E.C., Hall, S.A., & Shibli, S. (2019). *Project Management*. Sport Facility Operations Management.
13. Silva, D.N., Simões, A., Cardoso, C., Oliveira, D.E., Rittmeyer, J.N., Wehmuth, K., Lustosa, H., Pereira, R.S., Souto, Y.M., Vignoli, L.E., Salles, R., C., H.D., Ziviani, A., Ogasawara, E.S., Delicato, F.C., Pires, P.F., Pinto, H.L., Maia, L., & Porto, F.A. (2019). *A Conceptual Vision Toward the Management of Machine Learning Models*. ER Forum/Posters/Demos.
14. Singh, H. (2020). *Making an End-to-End Pipeline in SageMaker*.
15. Sun, C., Azari, N., & Turakhia, C. (2020). *Gallery: A Machine Learning Model Management System at Uber*. EDBT.
16. Weber, C., Hirmer, P., & Reimann, P. (2020). *A Model Management Platform for Industry 4.0 - Enabling Management of Machine Learning Models in Manufacturing Environments*. BIS.
17. Willumsen, P.L., Oehmen, J., Stingl, V., & Geraldini, J.G. (2019). *Value creation through project risk management*. International Journal of Project Management.
18. Xin, D., Miao, H., Parameswaran, A.G., & Polyzotis, N. (2021). *Production Machine Learning Pipelines: Empirical Analysis and Optimization Opportunities*. Proceedings of the 2021 International Conference on Management of Data.
19. Ye, W., Hu, R., & Enev, M. (2020). *Put Deep Learning to Work: Accelerate Deep Learning through Amazon SageMaker and ML Services*. Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.