МОЖЛИВОСТІ DATA SCIENCE* В ЦЕНТРАЛЬНИХ БАНКАХ: ОГЛЯД

ДМИТРО КРУКОВЕЦЬ[a]

[a]Національний банк України, Київ, Україна
E-mail: dmytro.krukovets@bank.gov.ua

Анотація У статті розглянуто основні напрями використання алгоритмів Data Science у центральних банках. Стаття містить огляд випадків використання Data Science, у тому числі для прогнозування макроекономічних і фінансових змінних, аналізу текстів (із газет, соціальних мереж та різних видів звітів), а також інших методів, які базуються або пов'язані з великими обсягами даних. Кожен із них до певної міри важливий для центральних банків загалом і Національного банку України зокрема. Їх застосовують для поліпшення формування стратегії політики, підвищення спроможності прогнозування і для інших цілей. Стаття сприятиме визначенню вектора дослідження у цій сфері, а також продемонструє, що популярність методів Data Science серед центральних банків із часом зростає. Крім того, у статті приділено увагу огляду напрацювань Національного банку України у цій сфері.

Класифікація JEL C45, C53, C82, E27, E37

Ключові слова Data Science, машинне навчання, обробка природної мови, макроекономіка, прогнозування

1. ВСТУП

Методи Data Science – інноваційний спосіб вирішення традиційних проблем центральних банків. Це широке поняття, яке поєднує машинне навчання та обробку даних. Перше – набір інструментів, які вчаться за допомогою доступних даних, розуміють закономірності і взаємодію між рядами і величинами. Вони можуть помітити зв'язки навіть тоді, коли люди не спроможні це робити (через великі обсяги даних і складність зв'язків). Друге означає можливий набір дій із самими даними: збір, перетворення, підготовка і візуалізація. На відміну від економетрики, яка зосереджена на вирішенні задач із нелінійним зміщенням шляхом зведення їх до лінійної форми, Data Science розширює можливості роботи з нелінійними зв'язками у системі. Ще одна відмінність полягає в тому, що економетрика основну увагу приділяє надійності методів, тоді як алгоритми машинного навчання набувають поширення через їхні відмінні результати[1]. До основних недоліків просунутих методів Data Science належить обмежена можливість інтерпретувати отримані результати. Саме тому використання цих методів не завжди можливе, оскільки від працівників центральних банків, які є спеціалістами в інших сферах, часто вимагають пояснення результатів (Kuhn & Johnson, 2013).

Інтерес до методів Data Science повернувся на початку 2010 року, коли були створені високоякісні моделі розпізнавання зображень, обчислювальна здатність збільшилася до достатнього рівня, а люди в багатьох галузях зрозуміли потенціал такого підходу. Саме тому в нашій статті особливу увагу буде приділено інструментам, які раніше рідко використовувалися в економіці і фінансовому секторі.

Нові елементи й інструменти починають проникати в дослідницьку діяльність і звичайні процеси в центральних банках. Удосконалені прогнозні моделі, що базуються на високочастотних даних, підвищують точність прогнозування, яке здійснюється за допомогою наявного набору інструментів, і можуть використовуватися як непоганий додаток або навіть замінник чинних моделей. Прогнозування не є єдиною можливою сферою застосування для алгоритмів Data Science. Методи обробки природної мови є ще однією сферою застосування алгоритмів Data Science для аналізу текстів. Вони здатні забезпечити аналіз тексту (новин, тексту із соціальних мереж) з метою оцінки реакції громадськості на політику і дії центрального банку або в роботі з пояснювальних досліджень.

Мета статті – детально розкрити застосування методів Data Science у центральних банках і навести приклади можливих випадків використання цих методів. Особливу увагу буде приділено дослідницькій діяльності, моделюванню і прогнозуванню. Автор не заглиблюватиметься в популярні технології використання великих даних: веб-скрепінг, а також наглядові і регуляторні технології (SupTech та RegTech). Причина полягає в тому, що в цих технологіях основна увага приділяється технічній реалізації та інформатиці, а не статистиці (економетриці) і математиці.

* В українськомовній літературі вживається також "Наука про дані".

[1] "Показовим прикладом є успіх алгоритму XGBoost, який став успішним завдяки його перемозі на декількох змаганнях із машинного навчання, а не завдяки його математичній обґрунтованості". Цитата зі статті, яка доступна за цим посиланням: https://towardsdatascience.com/from-econometrics-to-machine-learning-ee182f3a45d7.

Стаття має таку структуру. У розділі 2 розглянуто мотиви використання інструментів та алгоритмів Data Science у центральних банках. Розділ 3 містить огляд прогнозних моделей і підходів у рамках Data Science, розділ 4 – аналіз текстів, також наведено багато прикладів випадків використання цієї технології в центральних банках. Для повноти картини в розділі 5 розглянуто інші допоміжні методи. Насамкінець, у розділі 6 подано перелік напрацювань у цій сфері дослідників Національного банку, а в розділі 7 – стислий виклад висновків. Короткий опис визначень Data Science, які використовуються в зазначених працях, наведено в додатку.

2. ОГЛЯД ІНСТРУМЕНТІВ DATA SCIENCE, ЯКІ ЗАСТОСОВУЮТЬСЯ В ЦЕНТРАЛЬНИХ БАНКАХ

Центральні банки зацікавлені в застосуванні інструментів Data Science з кількох причин. Перша причина полягає в їхній новизні і можливості отримання точніших результатів, друга – у тому, що такі моделі можна ефективно застосовувати до мікрорівневих даних, "багатих" і деталізованих даних. Крім того, інструменти Data Science дають змогу використовувати інноваційні джерела інформації, які раніше мало використовувалися (наприклад, тексти), для отримання кращої попередньої оцінки настроїв населення та його очікувань щодо економіки, дій центрального банку тощо. Насамкінець, у більших обсягах даних міститься не менше інформації; питання лише в тому, як її добути.

Існує чимало першопроходців у сфері впровадження машинного навчання в центральних банках, які торували шлях для подальших досліджень у цій сфері. Це, наприклад, центральні банки Англії, Канади, Польщі та Індонезії. Однак у переважній більшості банків не видно ніяких ознак застосування цих методів, технології Data Science перебувають там у зародковому стані.

Більшість матеріалів про інструменти Data Science у сфері діяльності центральних банків подана лише у формі оглядів. Типовим прикладом є презентація, підготовлена представником Банку Англії Полом Робінсоном у 2018 році. У ній обговорюються проблеми, пов'язані з формуванням політики, а саме: неточні розрахунки, занадто складні моделі і недосконалі теорії, які лежать в основі таких моделей, а також внутрішні розбіжності. У цілому підходи, які базуються на використанні Big Data, і відповідні методи можуть сприяти вирішенню всіх цих проблем.

У презентації показано можливості використання машинного навчання як інструменту, який доповнює традиційні моделі. Наприклад, показники щодо ринку праці апроксимувалися та прогнозувалися на основі реклами вакансій та плину даних щодо пошуку роботи, які завантажувалися за допомогою інструментів веб-скрепінгу. Незважаючи на такі переконливі результати, у цих статистичних даних беруться до уваги лише користувачі, які є споживачами послуг компанії Google, і не враховуються споживачі послуг компаній Bing, Yahoo або тих, хто взагалі не користується інтернетом. Але навіть за таких обставин ця підвибірка має достатній рівень репрезентативності і може бути скоригована за допомогою базових агрегованих статистичних даних щодо користувачів послуг компанії Google.

На жаль, разом із новими підходами з'являються і нові ризики. По-перше, великий обсяг даних не обов'язково означає великий обсяг нової інформації. Деякі події, такі як високий рівень інфляції, пастки ліквідності, фінансова нестабільність і банкрутства банків, трапляються досить рідко. Частково це пов'язано з тим, що дії центральних банків спрямовані на недопущення настання таких подій. Тому не дивно, що обсяги інформації про такі події мізерні незалежно від обсягів наявних даних.

По-друге, моделі Data Science є переважно "чорними скриньками", зміст яких дуже важко пояснити, а це може бути неприйнятним для центральних банків, на відміну від IT-компаній, які регулярно отримують вигоду від використання таких моделей. Нарешті, більша деталізація даних підвищує ймовірність витоку конфіденційної інформації, що потребує встановлення додаткових вимог до рівня безпеки.

У робочому документі Банку Англії, розробленому Chakraborty & Joseph у 2017 році, наведено технічні деталі і випадки реального використання. У цьому 90-сторінковому документі за допомогою простих слів і формул зроблено огляд перетворення даних, інструментів оцінки і сучасних методів машинного навчання: метод наївного байєсу (Naive Bayes), метод k-найближчих сусідів [k-Nearest-Neighbours (k-NN)], нейронна мережа (Neural Network), метод опорних векторів [Support Vector Machine (SVM)], метод k-середніх (K-means) тощо. У цьому документі також ідеться про впровадження політики і випадки використання.

Перший приклад – це прогноз результатів банківського нагляду, який є класичною вправою з виявлення аномальних значень. У цій вправі модель "тренують" розпізнавати аномальну поведінку і виявляти статистичні викиди (аномальні значення). З метою побудови надійної моделі автори видалили частину даних, які були використані для створення цільової змінної. За таких умов випадковий ліс [Random Forest (RF)] виявляється найкращою моделлю після здійснення всіх оцінок (на предмет точності, прецизійності, повноти й оцінки F1).

Другий випадок – це прогнозування інфляції у Великобританії, одна з яскраво виражених класичних задач у роботі центральних банків. У більшості випадків алгоритми Data Science дають змогу отримати кращі результати, ніж традиційні економетричні моделі, причому найкращий результат отримуємо шляхом поєднання NN та SVM. Незважаючи на це, такі моделі, наголошують автори, є витратними з точки зору обчислень і потребують величезних обсягів даних, що не дуже часто трапляється в макроекономіці.

В останньому прикладі у статті йдеться про компанії-"єдинороги" у сфері фінансових технологій. Маються на увазі провідні, орієнтовані на технології фірми, які змінюють правила гри для всього сектору. Прикладом таких компаній є Uber (послуги таксі), AirBnB (готельний бізнес) і Glovo (послуги доставки). Їхня діяльність вплинула на сектор, у якому вони працюють, і певною мірою – на всю економіку. Автори побудували кластерну модель, яка ґрунтується на базі даних CrunchBase[2], і отримали кластер, що складається переважно з "єдинорогів". Однак навіть у цьому конкретному кластері

[2] На платформах, які містять інформацію про бізнес і приватні компанії, створено велику комплексну базу даних. https://data.crunchbase.com/docs/getting-started

багато компаній, які не є "єдинорогами". Таким чином, хоча ця модель і допомагає зрозуміти, виконання яких умов необхідне для досягнення компанією успіху, її не можна вважати достатньою.

На думку Per Nymand-Andersen (2017), радника Європейського центрального банку (ЄЦБ), "еволюція сервісів даних" створює велику нову сферу можливостей. Центральні банки не повинні втрачати таку можливість. Нині потоки спроможні відобразити стан економіки практично в режимі реального часу. Цю інформацію можна використати для розробки короткострокових стратегій, а ці стратегії потім – скоригувати на довгострокову перспективу. Фінансові установи змушені використовувати більші обсяги мікрорівневих даних з метою забезпечення своєї конкурентоспроможності. Фінансовий регулятор може також користуватися цими даними для кращого розуміння поведінки фінансових агентів та ефективного виконання наглядових функцій.

Підсумовуючи викладене, зазначимо, що інструменти Data Science можуть ефективно використовуватися додатково до наявних методів. Ці інструменти відкривають нові можливості у сфері прогнозування, аналізу й обробки даних. У наступних розділах ми розглянемо ці алгоритми детальніше, а також наведемо більше прикладів їхнього використання.

3. МАКРОЕКОНОМІЧНЕ ПРОГНОЗУВАННЯ І МОДЕЛЮВАННЯ

Упродовж минулих десятиліть дослідники в галузі макроекономіки побудували чимало економетричних моделей. Із часом ці моделі ставали складнішими, що давало змогу уникати різних зміщень під час оцінки. Незважаючи на це, дослідники почали використовувати сучасні алгоритми машинного навчання, які відіграють роль перспективних додаткових або альтернативних алгоритмів. Такі алгоритми є більш вимогливими з точки зору даних, але вони стають простішими у використанні завдяки поточним тенденціям світового розвитку.

Ми почнемо з прогнозування інфляції, яке є одним із найактуальніших завдань у центральних банках, особливо нині, коли в світі набуло популярності таргетування інфляції[3]. Праця Nakamura (2005) належить до найбільш ранніх досліджень на цю тему в даному огляді. У ті часи нейронні мережі не були популярним інструментом, тому праць, присвячених їм, дуже мало. Автор використовує квартальні дані з 1960-го до 2003 року. Його нейронна мережа проста і містить лише дві пари рівнянь з однією змінною, з'єднаних послідовно. Метод пошуку найкращих коефіцієнтів суттєво відрізняється від сучасного тим, що за старим методом береться сотня випадкових початкових значень, із яких обираються найкращі коефіцієнти, тоді як у новому методі використовується зворотне поширення оптимізованого рішення. Навіть за такого примітивного підходу нейронна мережа дала змогу отримати кращий результат, ніж авторегресійна бенчмарк-модель [AutoRegressive (AR) benchmark] на прогнозному горизонті від одного до чотирьох кварталів. Такі результати були отримані завдяки здатності нейронних мереж відображати нелінійні зв'язки. Основний висновок із цього полягає в тому, що нейронні мережі можуть ефективно використовуватися додатково до прогнозних моделей, які використовуються нині.

Складність мереж і методів зростає зі зростанням популярності галузі. Choudhary & Haider (2012) демонструють результати, отримані за допомогою нейронних мереж, на прикладі рядів даних щодо різних країн і порівнюють їх із результатами авторегресійної моделі [AR(1)]. Ця праця має складнішу структуру, ніж публікація, про яку йшлося раніше: у ній ідеться про дві мережі, які називаються гібридною і динамічною мережами [остання ближче до нейронної мережі зі зворотними зв'язками (RNN), але простіша], а також два типи поєднання цих мереж. У результаті нейронні мережі в більшості випадків дають змогу отримати кращі результати для короткострокового прогнозного горизонту, ніж авторегресійна модель [AR(1)] на основі бази даних щодо місячного рівня інфляції з липня 1991-го до червня 2008 року для 28 країн ОЕСР. Автор стверджує, що з огляду на нестабільність результатів порівняння економетричних та інших моделей протягом тривалого періоду часу є кращою стратегією. Отже, розробка інструментарію, до якого входить широкий набір інструментів, добре узгоджується зі стратегією.

Тим не менше складність моделі є не єдиною відмінністю. Світ перетворюється на "землю великих даних", у якій якість та обсяги даних збільшуються. Це впливає на процес прогнозування і прогнозні моделі. Medeiros et al. (2018) використали дуже велику базу місячних даних, яка називається FRED-MD і містить сотні рядів, що дають змогу прогнозувати інфляцію у США. У документі надано огляд деяких моделей – від бенчмарк-моделей і традиційних економетричних моделей до моделей Data Science. По-перше, алгоритми ледь здатні помітити нелінійні фактори, які можуть помітити ML-моделі, наприклад, такі як зв'язок між інфляцією і безробіттям. Зокрема, метод випадкового лісу [Random Forest (RF)] допоміг отримати найкращі результати для більшості прогнозних горизонтів, та й регресії Ridge і Lasso дали змогу отримати непогані результати. У процесі використання більшості моделей отримано проміжний продукт: перелік ознак, які були відібрані як найважливіші для пояснення розходжень у значеннях на кожному прогнозному горизонті. Результати, отримані за допомогою різних моделей, виявилися досить різними. Відповідно до результатів, отриманих шляхом застосування регресії Lasso, випуск і ціни є важливими змінними для пояснення інфляції. У свою чергу, згідно з результатами, отриманими за допомогою методу випадкового лісу і регресії Ridge, такими змінними є зайнятість, ціни і відсоткові ставки. Існує багато можливостей для аналізу і порівняння результатів, отриманих у результаті застосування різних моделей. З огляду на це завжди доцільно розширювати кількість моделей, що використовуються, навіть якщо в цих моделях застосовується однаковий набір змінних, оскільки це дає змогу розглянути питання з різних точок зору.

Jung et al. (2018) наводять докладний приклад застосування кількох методів машинного навчання, а саме – методу гнучкої сітки, алгоритму Super Learner та нейронної мережі зі зворотними зв'язками (RNN) для прогнозування зростання ВВП у кількох репрезентативних країнах. Особливо цікаво порівняти

[3] Відповідно до Щорічного звіту МВФ про валютні режими і валютні обмеження за 2018 рік.

результати, отримані за допомогою цих моделей, з офіційними прогнозами, які опубліковані у звіті "Перспективи розвитку світової економіки" (WEO), і які базуються на більш традиційних моделях. Метод гнучкої сітки та алгоритм Super Learner забезпечують вищий ступінь точності (вищий за еталонний показник на 35–80 відсотків) для прогнозного горизонту один квартал. Однак на прогнозному горизонті тривалістю один рік наведені вище методи мають набагато нижчу точність прогнозування [застосування нейронної мережі зі зворотними зв'язками (RNN) дало змогу отримати кращі результати для США, Великобританії і Німеччини, тоді як прогноз, опублікований у звіті "Перспективи розвитку світової економіки" (WEO), виявився точнішим для Іспанії, Мексики і В'єтнаму]. Ці алгоритми добре підходять для складання короткострокових прогнозів, а також можуть бути корисними для довгострокового прогнозування.

Наукастинг – це метод, який дає змогу прогнозувати значення змінних станом на сьогодні, і який використовується для прогнозування змінних, таких як ВВП, що публікуються зі значним розривом у часі. У відповідній літературі є багато прикладів застосування цього методу. Наприклад, він висвітлюється в праці Richardson et al. за 2019 рік (вона базується на їхньому дослідженні 2018 року, яке було допрацьоване і доповнене новими даними). Перелік наявних інструментів дуже широкий. До нього входять регресії Ridge та Lasso, байєсівська векторна авторегресія (Bayesian VAR), нейронні мережі, алгоритми бустингу, метод опорних векторів (SVM) та метод k-найближчих сусідів (k-NN). Масив даних також великий: внутрішня і міжнародна статистика, опитування і фінансові дані. Періодичність даних – від щоденних до щоквартальних. У цій праці йдеться про широкі можливості наукастингу ВВП. Дана методологія може також застосовуватися й щодо інших макроекономічних змінних.

Ще один приклад наукастингу ВВП наведено Bolhuis & Rayner у 2020 році. Масив даних налічує сотні змінних, які характеризують економіку Туреччини, деякі з них представлені кілька разів із різними перетвореннями. Це класичні змінні, які стосуються показників безробіття або поточного рахунку, і які поєднуються з різними індексами впевненості, а також зі змінними, які базуються на опитуваннях. Цей метод є поєднанням кількох стандартних методів машинного навчання, а саме: метод опорних векторів (SVM), метод градієнтного бустингу (GBM) та випадковий ліс (RF). Автори стверджують, що ці моделі доповнюють одна одну, а поєднання наукастів, побудованих за допомогою зазначених моделей, зменшує розмір похибки. Автори використовують різні комбінації: з однаковими вагами, а також такі, які базуються на відносній середньоквадратичній похибці (RMSE) окремих моделей. Результатом є повне перевищення ефективності поєднаних методів порівняно з окремими методами і бенчмарк-моделлю (традиційно для наукастингу ВВП використовується динамічна факторна модель).

Методи машинного навчання мають широку сферу застосування, яка не обмежується класичним прогнозуванням інфляції або макроекономічних показників. Gogas et al. (2014) пояснюють розрив випуску та інфляційний розрив за допомогою різних кривих дохідності. Модель, яка базується на методі опорних векторів (SVM), дає змогу змоделювати майбутні відхилення, що сприяє швидкому і належному реагуванню інструментами політики. Продовженням попередньої праці є дослідження, в якому здійснюється оцінка моделі, що базується на методі опорних векторів, за допомогою місячних даних щодо індексу Eurocoin Index і грошових агрегатів (M1, M2, M3). Це дослідження проводилося Gogas et al. у 2019 році. Автори доводять, що дані щодо грошової маси можуть використовуватися з метою прогнозування економічної активності в країнах зони євро. Це означає існування суттєвої залежності між економічною діяльністю й обсягом грошової маси. Гіпотеза щодо ефективності монетарної політики не може бути спростованою для країн зони євро.

Центральні банки займаються не лише макроекономічними питаннями, а й питаннями забезпечення фінансової стабільності. Petropoulos et al. (2018) розглянули питання застосування жорсткіших наглядових заходів з боку регуляторів. Регуляторам завжди доводиться балансувати між нижчим рівнем обмежень і забезпеченням фінансової стабільності. З огляду на це алгоритми для аналізу і запобігання різним типам ризиків нині користуються значним попитом. Автори створили великий, об'єднаний масив піврічних даних, який містить дані щодо кредитування більш як за 10 років і 354 часові ряди фінансових коефіцієнтів та макроекономічних змінних. Відношення кількості змінних і спостережень створює "прокляття розмірності" – тобто погану генералізацію. Дану проблему можна вирішити шляхом зменшення кількості змінних за допомогою алгоритму Boruta. Цей алгоритм, який базується на методі випадкового лісу (Random Forest), може бути використаний з метою відбору 65 найважливіших змінних. Відтак автори застосували метод eXtreme Gradient Boosting (XGBoost) та метод глибокої нейронної мережі (DNN) і порівняли результати, отримані за допомогою цих методів, із результатами, отриманими за допомогою латентного аналізу Діріхлє [Latent Dirichlet Analysis (LDA)] і моделей логістичної регресії (logit model). Метод XGBoost дав змогу отримати найкращі результати (вимірювалися за допомогою площі під кривою помилок AUROC), а також створити таблицю важливості змінних. Найважливішими були результати щодо акціонерного капіталу, наявності оборотного капіталу і покриття процентних витрат. Такі результати можуть застосовуватися як для бальної оцінки ризиків, так і для здійснення подальшого аналізу у цій сфері.

Скоринг у фінансовому секторі є одним із найперспективніших шляхів використання інструментів Data Science. Причини цього включають наявність великих обсягів високочастотних даних, складна структура даних, альтернативні джерела і види даних, які може бути складно використати у класичних економетричних моделях. Bazarbash (2019) у своїй праці представив розгорнуту дискусію щодо переваг та недоліків застосування алгоритмів машинного навчання для бальної оцінки платоспроможності позичальників, особливо у країнах із ринками, що розвиваються, які мають слабкі фінансові установи, а також доповнив її загальним оглядом методології. Автори приділяють особливу увагу перевагам методів Data Science для здійснення бальної оцінки платоспроможності позичальників: незначні витрати на здійснення досить точного аналізу позичальника невеликих сум у випадках,

коли залучення фінансового аналітика нерентабельне; здатність обробляти некількісну інформацію[4] і надавати їй більш кількісного характеру; здатність помічати нелінійні фактори і зменшувати ступінь інформаційної асиметрії. До недоліків належать проблеми, пов'язані з конфіденційністю, етичні проблеми, а також класичні проблеми керованих даними моделей, такі як погана реакція цих моделей на структурні розриви.

Праці, які вважаються спірними, можуть бути корисними лише у випадку підтвердження зроблених у них висновків технічними дослідженнями, в яких наведені вимірювані результати. Munkhdalai et al. (2019) протестували ряд моделей Data Science з метою побудови системи бальної оцінки платоспроможності позичальників та порівняння результатів цієї системи з результатами оціночної системи рейтингу позичальників за їхньою платоспроможністю (FICO). У процесі побудови системи автори використали автоматизований сітковий пошук з метою виявлення гіперпараметрів DS-моделей (застосовували повний перебір за кількома опціями для всіх алгоритмів) і алгоритми конструювання певних ознак (з метою зменшення їхньої кількості, а також з метою уникнення проблем надмірного навчання). Як і очікувалося, результат, отриманий за допомогою такої системи, був кращим від результату бенчмарк-систем.

Інша сфера застосування методів Data Science на фінансовому ринку – це забезпечення підтримки процесу прийняття рішень на кредитному ринку. Arora et al. (2019) пишуть про горизонтальний або вертикальний поділ під час створення портфеля. Ці стратегії дають змогу здійснювати належне управління ліквідністю, що важливо з точки зору ризику значних погашень, які можуть дестабілізувати фінансовий сектор. Одним з інструментів є здатність кількісно вимірювати вплив продажу на стан ринку. Метод випадкового лісу вдало справляється із цим завданням, прогнозуючи реакцію ринку порівняно краще, ніж традиційні моделі.

Остання праця, яку ми розглянемо в цьому розділі, присвячена досить незвичному виду прогнозування, що досліджувався Hatko у 2017 році. У статті розглядається відсутність відповідей в опитуваннях підприємств. Такі позиції зазвичай ігнорують або замінюють фіктивними змінними. Інший підхід – це прогнозування такої змінної, що базується на всіх відповідях інших підприємств, які мають подібні характеристики, і відповіді яких подібні до відповідей цієї фірми. Автор поділив цілу проблему на кілька підпроблем: підгрупа без відповідей (коли відповіді не надавалися зовсім) і підгрупа відповідей (коли відповіді не надавалися лише на окремі запитання). Перша була вирішена шляхом моделювання ймовірностей відповіді за допомогою поєднання методу логістичної регресії і методу кластеризації k-середніх. Друга – за допомогою застосування методу градієнтного бустингу (GBM) та алгоритму XGBoost до баз даних, у яких не вистачало даних, та до їхніх оціночних значень. Якість було оцінено шляхом використання кількох методів на основі середнього арифметичного (перехресної ентропії) та прогнозування за межі вибірки.

Необхідно приділити особливу увагу повністю сформованим комплексам, що використовують кілька інструментів, із тих, які описано вище. Вони мають простий інтерфейс користувача і здатні забезпечити комплексне вирішення конкретних проблем, перед якими постають центральні банки. Проєкт компанії Mindbridge[5] (Mindbridge.ai project) отримав щорічну нагороду Central Banking FinTech RegTech Global Awards як найкраще рішення у сфері машинного навчання. Їхній продукт дає змогу здійснити аналіз інформації про суб'єктів господарювання і будує шкалу ризиків, які є індикаторами ймовірності того, що якийсь конкретний суб'єкт виявиться шахраєм. Крім того, вони створили пошукову систему, яка допомагає регулятору здійснювати пошук і порівняння з іншими регуляторами, а також візуалізувати різні характерні риси суб'єктів господарювання. Після цього вони приєдналися до проєкту Банку Англії "Акселератор фінтеху" (Fintech Accelerator) і 2018 року отримали нагороду Central Banking Award за найкращу інновацію, що зайвий раз підкреслює відкритість Банку Англії до нових технологій, таких як алгоритми машинного навчання. Цей проєкт підтримує подальшу діджиталізацію наглядових процесів з метою підвищення якості і швидкості перевірок шляхом поєднання можливостей людини і штучного інтелекту.

Хоча в більшості випадків можуть використовуватися традиційні економетричні моделі, підходи Data Science дадуть змогу забезпечити додаткову точність. У більшості праць стверджується, що нові моделі можуть успішно використовуватися додатково до наявного інструментарію. Повний потенціал методів Data Science буде розкрито в наступному розділі у сфері, в якій не існує адекватних альтернативних методів.

На завершення цього розділу варто ще раз назвати сфери застосування алгоритмів Data Science: 1) прогнозування важливих макроекономічних змінних, таких як інфляція, ВВП, рівень безробіття тощо з можливим використанням альтернативних баз даних; 2) здійснення аналізу можливості прогнозування деяких змінних за допомогою інших змінних (тобто вплив другої змінної на першу змінну); 3) побудова різних індексів для прийняття рішень; 4) знаходження альтернатив експертним судженням у сферах, у яких класичні економетричні моделі неспроможні охопити і використати всю наявну інформацію; 5) заповнення прогалин і неспостережних змінних у даних.

4. АНАЛІЗ ТЕКСТУ

З розвитком Data Science центральні банки отримали можливість використовувати альтернативні джерела інформації. Наприклад, центральний банк Сінгапуру переглядає та фільтрує новини для виявлення подій, які потребують подальшої уваги (сповіщення). David R. Hardoon торкається цієї проблематики у своїй презентації на форумі Комітету Ірвінга Фішера з питань статистики центральних банків (IFC) у 2018 році. Таку роботу можуть виконувати аналітики, переглядаючи великі масиви тексту і відкидаючи неактуальні новини. Однак це досить рутинна робота. Водночас машинні алгоритми можуть узяти на себе цю роль, виконуючи такі задачі швидше та заощаджуючи час аналітиків.

Непередбачувані людські настрої становлять одне з основних джерел похибки в сучасних макроекономічних моделях. Ця проблема загалом вирішується шляхом

[4] Детальніший опис інформації, яку можна легко перевести у числовий формат (hard information), і інформації, яку складно перевести у цифровий формат (soft information), наведено в праці Liberti & Petersen, 2019.
[5] Загальний опис цього наведено на сайті Medium: https://medium.com/reciprocal-ventures/mindbridge-analytics-why-we-invested-9cdb2099ba

використання агрегованих даних та припущення, що суб'єкти діють здебільшого раціонально, тому впливом поведінкових ефектів поведінки можна знехтувати. Такий вплив можна також приблизно оцінити за допомогою даних із новин, соціальних мереж та інших текстових джерел. Цей розділ зосереджений на методах, які допомагають вирішити дану проблему.

Обробка природної мови [англійською – Natural Language Processing (звідси скорочення – NLP)] є галуззю науки Data Science, що вивчає алгоритми, які можуть "розуміти" текст, а не просто збирати статистику про нього. Це широка галузь знань, що передбачає кілька етапів та інструментів підготовки й обробки даних, вибір яких залежить від задачі, поставленої перед моделлю. Bholat et al. (2015) розглядають основні методи цієї галузі, комплексно розкривають мотивацію для їхнього використання, а також причини, з яких їх недооцінюють. У більшості праць, що згадуються в цьому розділі, йдеться про ті самі методи, які описані в статті: побудова словника, LDA та інші.

Текстовий аналіз не є новітньою сферою дослідження, однак попередні спроби центральних банків автоматизувати роботу з текстом не мали особливого успіху. Hansen на форумі IFC у 2018 році у своїй презентації розповідає про історію цих спроб. У 2007–2011 роках інструменти полягали здебільшого в пошуку певних слів або словосполучень для визначення емоційного забарвлення статей. Недолік такого підходу полягає в тому, що фрази "ефективність центрального банку низька" та "у багатьох статтях зазначено, що ефективність центрального банку низька, однак насправді все навпаки", маючи протилежне значення, розглядаються як подібні. Завдяки кращій обчислювальній ефективності та вищому попиту сучасна література може запропонувати кращі методології, починаючи з передових словникових методів, що використовують розуміння психологічної суті, та продовжуючи методами LDA і RNN, які можуть уловити контекст усього тексту. Реальним прикладом використання новітньої методології є вплив тексту релізу щодо інфляційного звіту Банку Англії на ціни облігацій та взаємозв'язок між заявами ФРС і шоками Ромер та Ромера[6]. Адекватні результати (76% точності) дають змогу застосовувати модель для отримання додаткової інформації про економіку та її динаміку.

Новини є опосередкованим показником суспільної реакції. Таким чином, деякі методи покликані будувати динамічні ряди показників, які допомагають пояснити важливі макроекономічні змінні. У статті Nicholas Apergis та Ioannis Pragidis 2019 року автори створили індикатор на основі настроїв у новинах і використали його для прогнозування біржового прибутку через базу даних зі статтями та статистикою на основі слів. Модель EGARCH-X, класична для фінансових цілей, демонструє кращу ефективність за використання індикаторів, створених на основі новин.

Іншим прикладом є модель, яка дає змогу пояснити спреди за облігаціями за допомогою індикаторів на основі новин (як на місцевому, так і на світовому рівнях) через часові ряди даних. Про це пишуть Fulop та Kocsis у 2018 році. Основним результатом є модель із використанням новин, яка демонструє значне збільшення коефіцієнта детермінації порівняно з простою макроекономічною моделлю. Це один із найкращих прикладів, коли методи NLP та новинні дані є основним елементом, а не просто доповненням, що слугує для пояснення кількох додаткових відсоткових пунктів дисперсії.

Створення індикаторів для моделей є не єдиною метою інтелектуального аналізу новин. Rybinski (2019) продемонстрував модель, яка дає змогу аналізувати статті про Національний банк Польщі (НБП) у головній польській газеті "Rzeczpospolita" за двадцятирічний період. Інтелектуальний аналіз тексту допомагає оцінити зв'язок між актуальними темами в економіці (визначеними на основі бальної оцінки, що застосовується до новин) та оцінкою в результаті переговорів Комітету з монетарної політики (польською мовою – RPP). НБП є ключовим органом, який приймає рішення в кількох сферах (наприклад, інфляція або процентні ставки). Але не в усіх сферах (скажімо, сюди не входять сфери державних фінансів чи фіскального сектору). Модель демонструє, що перші теми висвітлюються ЗМІ більше в періоди діяльності Комітету з монетарної політики, а другі – ні.

"Слова – нові цифри: індекс збіжних індикаторів циклу ділової активності, отриманий із новин", – назва праці, написаної Thorsrud у 2016 році, говорить сама за себе. Основною ідеєю є апроксимація для циклу ділової активності США за допомогою моделі LDA. Автор досліджував теми ділових газет, їхню динаміку (показник актуальності теми певного дня) та квартальне зростання ВВП у змінній у часі динамічній факторній моделі.

У багатьох випадках для розробки стратегії центральним банкам необхідно оцінити очікування суб'єктів господарювання. Зазвичай для цих цілей використовують опитування. Однак вони досить дорогі, особливо якщо говорити про високоякісні опитування (робастність до багатьох факторів, розмір та однорідність вибірки тощо). Можливість оцінки очікувань за допомогою новин є гарним доповненням до набору інструментів центрального банку. Zulen та Wibisono (2018) запровадили модель, яка прогнозує очікування суспільства щодо змін облікової ставки на основі новин (за чотирма категоріями: відсутність інформації, відсутність змін, відсутність підвищення або відсутність скорочення) та порівняли результати з індексом опитування Bloomberg. Результати були цілком задовільними. Цей приклад класифікації дав до 84% точності в моделі XGBoost, що є прекрасним показником, ураховуючи незначні витрати і переваги у швидкості такого методу. Переваги мають вирішальне значення для наукастингу та кращого розуміння поточного економічного становища.

Моніторинг новин допомагає прогнозувати шоки, такі як збройні конфлікти. Незважаючи на вражаючу ефективність сучасних інструментів прогнозування, існує ймовірність настання не зафіксованих раніше шоків через їхню можливу незвичну чи рідкісну природу походження, несезонний характер тощо. Mueller і Rauh (2017) написали статтю щодо прогнозування політичного насильства, такого як збройні конфлікти чи

[6] Шоки Ромер та Ромера (шоки RR) описані у відповідній праці за 2004 рік.

громадянські війни. Автори брали національні новини, відбирали теми та за допомогою дослідження динаміки, різних байєсівських методів, вибірки Гіббса та багатьох інших інструментів оцінювали ймовірність конфлікту всередині країни. Результати були вражаючими: завдяки моделі вдалося передбачити збройний конфлікт, що наближався, з імовірністю 70% (при цьому результати були лише на 20% хибно-позитивними). Такі моделі можуть значно поліпшити якість результатів структурних економічних моделей і можуть бути використані для ефективнішої розробки сценаріїв.

Дані соціальних мереж є вагомим джерелом інформації на мікрорівні в режимі реального часу, яким можуть користуватися центральні банки. Однак зазначене джерело використовується недостатньо. Таким чином, ця інноваційна галузь дослідження може дати навіть кращі результати, ніж моделі, засновані на новинах. На сьогодні в цій сфері є кілька незавершених проєктів. Наприклад, Angelico et al. (2018) критично висловлюються з приводу частоти, з якою проводяться опитування (один раз на місяць або навіть рідше), що, в свою чергу, не дає змоги негайно аналізувати реакцію на певні події. А отже, вони пропонують використовувати відфільтровані та підготовлені дані з твіттера для оцінки інфляційних очікувань. Такі результати демонструють високу кореляцію з показниками, отриманими в результаті опитувань, та ринковими показниками; водночас їх можна отримати в режимі реального часу та з низькими витратами. Це може слугувати доповненням до традиційних методів оцінки уявлень.

Corea (2016) описує інший варіант – використання даних із твіттера для оцінки настроїв інвесторів на фондовому ринку. Наприклад, станом на 2016 рік було 88000 твітів щодо акцій Apple. Проаналізувавши їх, можна дати приблизну оцінку очікуванням населення, а отже, спрогнозувати його поведінку. На жаль, результати досить неоднозначні, що свідчить про необхідність використання комплексних моделей та ретельнішої підготовки даних.

Дані соціальних мереж є безперервними. Ця особливість дає змогу вловлювати настрої, які неможливо отримати на регулярній основі під час опитувань. Таким чином, їх можна використовувати не лише для прогнозування, а й для досліджень. Stiefel та Vives (2019) виявили значну залежність між очікуваннями щодо індексу інтервенції ЄЦБ та спредами за облігаціями. Такі дані дають змогу працювати з динамікою настроїв та з чутками, що майже неможливо за допомогою інших інструментів. Нарешті, дані соціальних мереж можна використовувати і як доповнення до моделей прогнозування, і як незалежний елемент для окремих досліджень.

Наприкінці розділу наводиться кілька праць щодо комунікації між різними суб'єктами економіки. Це органи державної влади, люди, які проживають у країні, міжнародні установи. Перша праця, написана Fayad et al. у 2020 році, дає певне уявлення про комунікацію між органами державної влади та МВФ під час консультацій відповідно до статті IV Статуту МВФ[7]. Звіти співробітників (близько 2600 звітів) становлять масив даних за період із 2000 до 2018 року, що також уможливлює аналіз динаміки такої комунікації. Для початку з кожного звіту було взято кілька "значущих" абзаців. За допомогою "словникового методу" їм присвоїли тему, що, у свою чергу, забезпечило точність на рівні 89% (порівняно з меншим набором даних, підготовленим уручну). Потім для виявлення настроїв у цих абзацах застосовувалася провідна техніка BERT, що використовується для цілей NLP. Точність результатів сягнула 81%. Техніка непогано спрацювала у випадках, коли органи влади погоджувалися чи не погоджувалися з радниками МВФ, однак у випадках змішаної відповіді така техніка не впоралася. Ці результати можуть бути корисними радникам МВФ для вдосконалення своєї програми та пошуку найефективніших векторів співпраці.

Аналіз внутрішніх комунікацій (наприклад, обговорення в рамках Комітету з монетарної політики) центрального банку може бути корисним для багатьох сфер, у тому числі для підвищення прозорості, що є одним із найважливіших факторів для центральних банків. Донедавна дослідження ефекту прозорості було обмежене використанням періодичних фіктивних змінних. На сьогодні можна відстежувати ефект безпосередньо через динаміку тем за відповідні періоди (до, під час та після переходу до більш прозорої поведінки). Стаття з цього питання, написана Hansen et al. (2018), підтверджує гіпотезу про позитивний ефект дисципліни та негативний ефект підпорядкування внаслідок підвищення прозорості, а також гіпотезу про структурні зміни. На основі стенограм зустрічей Федерального комітету відкритих ринків автори виявили значне підвищення обсягу комунікації між членами (обговорення тих самих тем) за наявності дисципліни (підготовка до зустрічей, що підвищує інформативність) та підпорядкування (уникнення відвертого висловлювання власної думки, що знижує рівень інформативності). Незважаючи на необхідність прозорості, основний висновок полягає в тому, що методи NLP роблять можливими дослідження у сферах, які досі були здебільшого недоступними.

Остання праця, яка висвітлюється в цьому розділі, написана Cedervall та Jansson у 2018 році. Вона тісно пов'язана з попередньою та є гарним прикладом аналізу динаміки тем. Автори звернули особливу увагу на "цінність для бізнесу" практики складання швидкого огляду звіту за допомогою машинних технік. Для успіху у світі доступних даних бізнесові необхідно першим розуміти дані та передвісники, а не отримувати повні звіти із затримкою. Тому бізнес гідно оцінить оперативну необроблену та стислу інформацію. Висновок такий: прозорість залежить не лише від якості одного комунікаційного потоку, а й від різноманітності цих потоків. Тоді як дехто може краще сприймати інформацію в тому чи іншому форматі, цей інструмент допомагає досягнути різноманітності за відносно низьких витрат.

На закінчення розділу підсумуємо, що інтелектуальний аналіз тексту є широкою галуззю дослідження, навіть якщо ми говоримо про аналіз новин, хоча мова не лише про них. Центральні банки досі широко не користуються аналізом інформації із соціальних мереж для прогнозування поведінки, що робить цю сферу перспективною для новітніх досліджень. У розділі було описано багато сценаріїв використання інтелектуального аналізу тексту, а саме: 1) вивчення новин для розробки індикатора

[7] Це серія консультацій з окремими представниками країн у рамках консультацій відповідно до статті IV Статуту МВФ. Тут можна знайти короткий опис: https://www.imf.org/external/about/econsurv.htm

з метою прогнозування різних макроекономічних та фінансових часових рядів, таких як спреди за облігаціями та біржовий прибуток; 2) моделювання індексу довіри, прозорості та інших соціальних взаємодій на основі новин; 3) прогнозування ймовірності шоків, які раніше неможливо було передбачити, але про які експерти могли здогадатися; 4) вивчення результатів проксі-опитування щодо очікувань для різних рядів, наприклад, інфляційних очікувань; 5) дослідження різних ефектів, побудованих на комунікації; 6) полегшення в деяких випадках рутини робочого процесу; 7) дослідження соціальних мереж, які є одним із найкращих осередків для розуміння індивідуальної поведінки та настроїв.

5. ІНШІ АЛГОРИТМИ

Data Science не обмежується вищезазначеними алгоритмами. Існує велика кількість методів на стику аналізу даних, статистики та IT. Серед технік, які широко застосовують дослідники, можна назвати веб-скрепінг. Інструмент бере дані в реальному часі безпосередньо з веб-сайтів. Відомий проєкт у макроекономічній сфері, в якому використовується веб-скрепінг, називається "Мільярди цін" ("Billion Prices Project"). У ньому інформація про ціни збирається з роздрібних веб-сайтів та використовується замість (або як доповнення до) офіційних даних про рівень цін. Про цей проєкт було написано багато праць, серед них Cavallo (2013). Автор використав отримані за проєктом дані, щоб поставити під сумнів достовірність офіційної інформації про інфляцію в Аргентині. Він об'єднав ряди в компоненти, подібні до офіційного кошика споживчих товарів, щоб частково продемонструвати інфляцію на прикладі товарів із кошика. У цьому випадку результати між зібраними та офіційними цінами можуть різнитися через розбіжності в методології та з інших причин. Однак праця продемонструвала: інфляція за зібраними даними вдвічі вища від офіційної, що є доказом маніпуляцій з офіційними даними.

Google Trends – ще один метод із використанням інтернету, який схожий по суті на веб-скрепінг, але використовуються дані пошукових запитів користувачів. Цей метод почав працювати у 2006 році, однак у класичному економічному середовищі його використання почали інтенсивно обговорювати 5–10 років по тому. Per Nymand-Andersen на форумі IFC у 2018 році продемонстрував модель прогнозування даних про реєстрацію автомобілів (які на офіційному рівні подаються із затримкою) за допомогою даних Google Trends (кількох пошукових запитів, пов'язаних із купівлею автомобіля). Була доведена гіпотеза, що якщо люди починають шукати можливість придбати машину, то багато хто з них зробить це вже найближчим часом.

Центробанківська спільнота нині активно обговорює інші методики, які стосуються, швидше, "великих даних". Розробка та підтримка бази даних є важливою частиною впровадження алгоритмів Data Science, про що також зазначив Erwin Rijanto у своїй вступній промові на форумі IFC 2018 року. У рамках того ж форуму Renaud Lacroix продовжує обговорення цього питання у своїй презентації про проєкт побудови мультидисциплінарної платформи деталізованих даних. Soramaki (2018) наводить гарний приклад популярних рішень Regtech та Suptech. Він описав продукт FNA Ltd. для аналізу, моніторингу та візуалізації трансакцій між компаніями та установами, їхньої взаємодії у фінансовому плані. Кілька наочних підходів та статистичні дані сприяють розумінню мережі (фінансової системи). Кілька сценаріїв використання стосувалися цін на житло безпосередньо до, під час та після світової фінансової кризи, виявлення шахрайства під час здійснення грошових переказів тощо. Однак зазначена праця лише побіжно торкається теми дослідження і не є предметом розгляду нашої статті.

Останній приклад використання великих даних менше стосується діяльності центральних банків, однак є дуже гарним прикладом комплексної структури в економіці, яку можна пояснити або оцінити за допомогою методів Data Science. У 2018 році Fan Liang et al. описали проєкт органів влади Китаю, що дає змогу оцінити "надійність" китайських громадян та організацій. Моделлю передбачається оцінка за численними критеріями: що саме люди купують у магазинах, де проводять час, чи вчасно сплачують борги, хто їхні друзі та багато іншого. За допомогою цих даних виставляється бальна оцінка, яка відображає ймовірність повернення кредиту. Відповідно існує система винагород і покарань для тих, хто має відповідно високу та низьку бальні оцінки в системі соціального кредиту. Можна отримати багато даних стосовно конкретної людини та інформації про взаємозв'язки між цими даними, щоб поставити таку оцінку, яка відображає поведінку в певній ситуації. Це унікальна технологія, яка може використовуватися центральними банками для розуміння суспільної думки та очікувань із надзвичайною точністю. Однак виникають питання щодо витрат та етичної складової. Висновком цієї праці та цього продукту є те, що більшість речей, навіть таких нестабільних, як людська поведінка, можна виміряти і спрогнозувати з великою точністю. Усе залежить від даних.

Підбиваючи підсумки, зазначимо, що існує безліч методів та інструментів, пов'язаних із цією темою, про які слід було б згадати, однак вони виходять за рамки безпосередніх завдань нашої статті.

6. ПРОЄКТИ НАЦІОНАЛЬНОГО БАНКУ УКРАЇНИ

Багато проєктів на базі DS перебувають на різних стадіях завершення в Національному банку України. Метою цього розділу є їхнє висвітлення.

Проєкт щодо офіційних щомісячних дезагрегованих даних про інфляцію має на меті знайти ступінь розбіжності між рядами та кластеризувати їх у кілька груп. Це поставить під сумнів доцільність поточного поділу загальної споживчої інфляції на чотири основні категорії, виявить ряд із найвищим ефектом перенесення обмінного курсу та дослідить відносні ціни між товарами, які торгуються, і товарами, які не торгуються.

Також використовуються кілька незначних додаткових моделей для прогнозування інфляції. Вони засновані на методах випадкового лісу (Random Forest) та GBM (зокрема, XGBoost), призначених для прогнозування компонентів базової інфляції. Подальші дослідження можливостей прогнозування все ще на стадії розробки.

Увага НБУ як органу нагляду, метою якого є сприяння фінансовій стабільності, зосереджена на ризиках у

кредитному секторі. Дослідники вивчають кредитні ризики в різних контекстах, починаючи від даних окремих агентів і закінчуючи агрегованими індексами за всіма клієнтами в цілому у конкретному банку. Логістична регресія слугує основним задокументованим еталонним методом. Однією з причин для цього є те, що такий підхід допомагає надати пояснення, коли існує необхідність аргументованої відмови у видачі кредиту. Стосовно додаткових інструментів ми можемо згадати статтю Pokidin 2015 року. Автор досліджує, наскільки метод SVM ефективніший від еталонного методу та кількох інших моделей у частині корпоративних кредитних ризиків. Відповідь – "несуттєво". Іншим прикладом є проєкт на базі XGBoost для виявлення шахрайства на рівні МСП.

У своїй статті 2016 року Rashkovan та Pokidin розглядають групи банків, які кластеризуються за кількома типами залежно від своєї бізнес-моделі кредитування. Автори досліджують їхню динаміку під час основних змін у банківському секторі протягом 2014–2016 років та роблять спробу визначити, які з них були найризикованішими. Для цього автори використали алгоритм розширення до нейронних мереж. Урешті-решт, було розроблено показник ризику для банківського сектору, який давав можливість визначити 92% банків, що не виконують зобов'язання. Таким чином, це інформативний та корисний продукт для регуляторних органів.

Для НБУ залишається відкритим питання оцінки суспільних настроїв. Незважаючи на численні різноманітні опитування, триває пошук ефективніших способів їхньої оцінки як із точки зору витрат, так і з точки зору точності отриманих даних. Отже, розробка алгоритму для аналізу новин перебуває на завершальній стадії. Проєкт складатиметься з моделі (та кількох дослідницьких праць на її основі) для вивчення сфери медіа-середовища в Україні під різними кутами.

Веб-скрепінг цін на товари з інтернет-магазинів – це "гарячий" проєкт, у рамках якого проводиться велика кількість досліджень та застосовуються моделі на основі даних. Одне з досліджень уже опублікували у 2018 році Faryna, Talavera та Yukhymenko. Вони побудували модель, яка об'єднує ці ряди у складові інфляційного кошика. Цей набір даних охоплює 46% від загального офіційного споживчого кошика, визначеного Держстатом. Метою статті було дослідити, наскільки ціни в інтернеті відповідають офіційній статистиці та які фактори впливають на ефективність відображення інфляції за допомогою веб-скрепінгу. Цей проєкт отримує подальший розвиток та набуває здатності вирішувати інші задачі, зокрема такі, як групування рядів, як у проєкті Abe, Shinozaki 2018 року. Подібна техніка зустрічається ще в кількох проєктах, присвячених ринку праці та його щільності.

Дослідники НБУ успішно працюють у цій сфері, беручи участь у найважливіших сферах, реалізуючи проєкти на різних стадіях завершення. Однак зробити ще необхідно дуже багато.

7. ВИСНОВОК

Кількість досліджень на цю тему, навіть ураховуючи стале зростання їхньої кількості протягом останніх років, залишає бажати кращого. Нижче подана цитата з короткого огляду до однієї з праць, представленої в попередньому розділі: "Важливо зазначити, що нестача публікацій у цій сфері пов'язана не з обмеженою застосовністю машинного навчання, а з тим, що це лише початковий етап його розвитку". Це було написано 11 січня 2018 року. Відтоді ситуація поліпшилася, але не настільки, як би хотілося.

Як і очікувалося, в роботі центральних банків найефективнішими та найчіткіше прописаними моделями у сфері Data Science є ті, що застосовуються в галузі прогнозування. Поки зростає якість та різноманітність даних, є можливість удосконалити набір інструментів прогнозування за рахунок моделей, які комплексно поєднують різні дезагреговані ряди даних у спосіб, подібний до "чорної скриньки". Найкраще, що можуть зробити підрозділи, які займаються моделюванням, – це інвестувати в цю сферу більше часу.

З іншого боку, аналіз тексту також знаходить своє застосування. Це досить складне завдання, особливо в країнах, де англійська мова не є основною мовою спілкування. Воно пов'язано з необхідністю мати високоякісний словниковий запас. Для розвинутих економік дана сфера цікава, оскільки є новим способом удосконалення поінформованості в економічних питаннях із кращим урахуванням поведінки людей. Однак для економік із ринками, що розвиваються, співвідношення витрат і переваг низьке, тому недоцільно витрачати на це багато часу.

Нові джерела даних (веб-скрепінг, Google Trends) є ефективним та недорогим доповненням до існуючого набору інструментів, який може дати більше можливостей для використання передових методів Data Science. Ще одна з причин пов'язана з поліпшенням комунікації з іншими центральними банками. Спільні проєкти можуть удосконалити якість та різноманітність даних. Умови для цього сприятливі. У цій сфері проводяться великі конференції та часті зустрічі.

Сподіваємося, наша стаття, в якій зроблено огляд найбільш багатообіцяючих векторів дослідження, сприятиме та заохочуватиме дослідницьку діяльність на стику сфери відповідальності центральних банків та можливостей Data Science.

ЛІТЕРАТУРА

Abe, N., Shinozaki, K. (2018). Compilation of experimental price indexes using Big Data and Machine Learning: A comparative analysis and validity verification. Bank of Japan Working Paper Series, No. 18-E-13. Bank of Japan. Retrieved from https://www.boj.or.jp/en/research/wps_rev/wps_2018/data/wp18e13.pdf

Angelico, C., Marcucci, J., Miccoli, M., Quarta, F. (2018). Can we measure inflation expectations using Twitter? Harnessing Big Data & Machine Learning Technologies for Central Banks (Rome, March 26). Retrieved from https://www.bancaditalia.it/pubblicazioni/altri-atti-convegni/2018-bigdata/Miccoli_Presentazione_Twitter_Workshop.pdf

Apergis, N., Pragidis, I. (2019). Stock price reactions to wire news from the European Central Bank: evidence from changes in the sentiment tone and international market indexes. International Advances in Economic Research, 25, 91–112. https://doi.org/10.1007/s11294-019-09721-y

Arora, R., Fan, C., Leblanc, G. (2019). Liquidity management of Canadian corporate bond mutual funds: A Machine Learning approach. Staff Analytical Note, 2019-7. Bank of Canada. Retrieved from https://www.bankofcanada.ca/2019/02/staff-analytical-note-2019-7/

Bazarbash, M. (2019). FinTech in financial inclusion Machine Learning applications in assessing credit risk. IMF Working Papers, WP/19/109. International Monetary Fund. Retrieved from https://www.imf.org/en/Publications/WP/Issues/2019/05/17/FinTech-in-Financial-Inclusion-Machine-Learning-Applications-in-Assessing-Credit-Risk-46883

Bholat, D., Hansen, S., Santos, P., Schonhardt-Bailey, C. (2015). Text mining for central banks. Centre for Central Banking Studies Publication. Bank of England. Retrieved from https://www.bankofengland.co.uk/-/media/boe/files/ccbs/resources/text-mining-for-central-banks.pdf

Bolhuis, M., Rayner, B. (2020). Deus ex machina? A framework for macro forecasting with Machine Learning. IMF Working Papers, WP/20/45. International Monetary Fund. Retrieved from https://www.imf.org/en/Publications/WP/Issues/2020/02/28/Deus-ex-Machina-A-Framework-for-Macro-Forecasting-with-Machine-Learning-49094

Cavallo, A. (2013). Online and official price indexes: Measuring Argentina's inflation. Journal of Monetary Economics, 60(2), 153-165. http://dx.doi.org/10.1016/j.jmoneco.2012.10.002

Cedervall, A., Jansson, D. (2018). Topic classification of Monetary Policy Minutes from the Swedish Central Bank. Examensarbete Inom Technology, Grundnivå, 15 Hp. Stockholm, Sverige. Retrieved from http://www.diva-portal.org/smash/get/diva2:1272108/FULLTEXT01.pdf

Chakraborty, C., Joseph, A. (2017). Machine Learning at central banks. Staff Working Paper, 674. Bank of England. Retrieved from https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2017/machine-learning-at-central-banks.pdf

Choudhary, A., Haider, A. (2012). Neural network models for inflation forecasting: an appraisal. Applied Economics, 44(20), 2631-2635. https://doi.org/10.1080/00036846.2011.566190

Corea, F. (2016). Can Twitter proxy the investors' sentiment? The case for the technology sector. Big Data Research, 4(C), 70–74. https://dl.acm.org/doi/10.5555/2991306.2991336

Faryna, O., Talavera, O., Yukhymenko, T. (2018). What drives the difference between online and official price indexes? Visnyk of the National Bank of Ukraine, 243, 21-32. https://doi.org/10.26531/vnbu2018.243.021

Fayad, G., Huang, C., Shibuya, Y., Zhao, P. (2020). How do member countries receive IMF policy advice: Results from a state-of-the-art sentiment index. IMF Working Papers, WP/20/7. International Monetary Fund. Retrieved from: https://www.imf.org/en/Publications/WP/Issues/2020/01/17/How-Do-Member-Countries-Receive-IMF-Policy-Advice-Results-from-a-State-of-the-art-Sentiment-48937

Fulop, A., Kocsis, Z. (2018). News-based indices on country fundamentals: do they help explain sovereign credit spread fluctuations. MNB Working Papers, 1. Magyar Nemzeti Bank. Retrieved from https://www.mnb.hu/letoltes/mnb-wp-2018-1-final-1.pdf

Gogas, P., Papadimitriou, T., Matthaiou, M., Chrysanthidou, E. (2014). Yield curve and recession forecasting in a Machine Learning framework. Computational Economics, 45, 635–645. https://doi.org/10.1007/s10614-014-9432-0

Gogas, P., Papadimitriou, T., Sofianos, E. (2019). Money neutrality, monetary aggregates, and machine learning. Algorithms, 12(7), 137. https://doi.org/10.3390/a12070137

Hansen, S. (2018). Measuring market and consumer sentiment and confidence. Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_21.pdf

Hansen, S., McMahon, M., Prat, A. (2018). Transparency and deliberation within the FOMC: a computational linguistics approach. The Quarterly Journal of Economics, 133(2), 801–870. https://doi.org/10.1093/qje/qjx045

Hardoon, D. (2018). Exploring big data to sharpen financial sector risk assessment. Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_28.pdf

Hatko, S. (2017). The Bank of Canada 2015 retailer survey on the cost of payment methods: nonresponse. Technical Report, No. 107. Bank of Canada. Retrieved from: https://www.bankofcanada.ca/wp-content/uploads/2017/03/tr107.pdf

Jung, J., Patnam, M., Ter-Martirosyan, A. (2018). An algorithmic crystal ball: forecasts-based on Machine Learning. IMF Working Papers, WP/20/7. International Monetary Fund. Retrieved from https://www.imf.org/en/Publications/WP/Issues/2018/11/01/An-Algorithmic-Crystal-Ball-Forecasts-based-on-Machine-Learning-46288

Kuhn, M., Johnson, K. (2013). Applied Predictive Modeling. Springer. https://doi.org/10.1007/978-1-4614-6849-3

Lacroix, R. (2018). The Bank of France datalake. Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_26.pdf

Liang, F., Das, V., Kostyuk, N., Hussain, M. (2018). Constructing a data-driven society: China's Social Credit System as a state surveillance infrastructure. Policy & Internet, 10(4), 415-453. https://doi.org/10.1002/poi3.183

Liberti, J., Petersen, M. (2019). Information: hard and soft. The Review of Corporate Finance Studies, 8(1), 1–41. https://doi.org/10.1093/rcfs/cfy009

Medeiros, M., Vasconcelos, G., Veiga, A., Zilberman, E. (2019). Forecasting inflation in a data-rich environment: the benefits of Machine Learning methods. Journal of Business & Economic Statistics. https://doi.org/10.1080/07350015.2019.1637745

Mueller, H., Rauh, C. (2017). Reading between the lines: prediction of political violence using newspaper text. American Political Science Review, 112(2), 358-375. https://doi.org/10.1017/S0003055417000570

Munkhdalai, L., Munkhdalai, T., Namsrai, O., Yun Lee, J., Ho Ryu, K. (2019). An empirical comparison of Machine Learning methods on bank client credit assessments. Sustainability, 11(3), 699. https://doi.org/10.3390/su11030699

Nakamura, E. (2005). Inflation forecasting using a neural network. Economics Letters, 86(3), 373-378. https://doi.org/10.1016/j.econlet.2004.09.003

Nymand-Andersen, P. (2017). Big data in central banks: Central Banking focus report. Central Banking Publications. Retrieved from https://www.centralbanking.com/media/download/24906/download

Nymand-Andersen, P. (2018). Google econometrics: nowcasting euro area car sales and big data quality requirements. Statistics Paper Series, No. 30. European Central Bank. Retrieved from https://www.ecb.europa.eu/pub/pdf/scpsps/ecb.sps30.en.pdf

Petropoulos A., Siakoulis V., Stavroulakis E., Klamargias A. (2018). A robust Machine Learning approach for credit risk analysis of large loan-level datasets using deep learning and extreme gradient boosting. Ninth IFC Conference on "Are post-crisis statistical initiatives completed?" (Basel, 30-31 August 2018). Retrieved from: https://www.bis.org/ifc/publ/ifcb49_49.pdf

Pokidin, D. (2015). National Bank of Ukraine econometric model for the assessment of banks' credit risk and support vector machine alternative. Visnyk of the National Bank of Ukraine, 234, 52-72. https://doi.org/10.26531/vnbu2015.234.052

Rashkovan, V., Pokidin, D. (2016). Ukrainian banks' business models clustering: application of Kohonen neural networks. Visnyk of the National Bank of Ukraine, 238, 13-38. https://doi.org/10.26531/vnbu2016.238.013

Richardson, A., Mulder, T., Vehbi, T. (2019). Nowcasting GDP using Machine Learning algorithms: A real-time assessment. Discussion Paper, 2019-03. Reserve Bank of New Zeland. Retrieved from https://www.rbnz.govt.nz/research-and-publications/discussion-papers/2019/dp2019-03

Rijanto, E. (2018). Opening remarks. International Seminar on Big Data "Building Pathways for Policy-Making with Big Data" (Bali, 26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_02.pdf

Robinson, P. (2018). Big data: new insights for economic policy – The Bank of England Experience. IFC – Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_12.pdf

Romer, C., Romer, D. (2004). A new measure of monetary shocks: Derivation and implications. NBER Working Paper Series, 9866. National Bureau Of Economic Research. Retrieved from https://www.nber.org/papers/w9866.pdf

Rybinski, K. (2019). A Machine Learning framework for automated analysis of central bank communication and media discourse. The case of Narodowy Bank Polski. Bank i Kredyt, 50(1), 1-20. Retrieved from http://bankikredyt.nbp.pl/content/2019/01/BIK_01_2019_01.pdf

Soramaki, K. (2018). Introduction to network science & visualization. IFC – Bank Indonesia International Workshop and Seminar on "Big Data for Central Bank Policies / Building Pathways for Policy Making with Big Data" (Bali, Indonesia, 23-26 July 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb50_10.pdf

Stiefel, M., Vives, R. (2019). 'Whatever it Takes' to change belief: evidence from Twitter. AMSE Working Papers, Nr. 07. Aix-Marseille School of Economics. Retrieved from https://www.amse-aixmarseille.fr/sites/default/files/working_papers/wp_2019_-_nr_07_0.pdf

Thorsrud, L. (2016). Words are the new numbers: A newsy coincident index of business cycles. Journal of Business & Economic Statistics, 38(2), 393-409. https://doi.org/10.1080/07350015.2018.1506344

Zulen, A., Wibisono, O. (2018). Measuring stakeholders' expectations for the central bank's policy rate. Ninth IFC Conference on "Are post-crisis statistical initiatives completed?" (Basel, 30-31 August 2018). Retrieved from https://www.bis.org/ifc/publ/ifcb49_50.pdf

ДОДАТОК

Короткий опис інструментів Data Science, що згадуються в зазначених працях

Еластична сітка, Elastic Net – метод регуляризації регресії. Відповідно до цього методу початкова задача мінімізації методом найменших квадратів (МНК) доповнюється добутком лямбда з коефіцієнтом бета та ще однієї лямбда з бета у квадраті. Це допомагає зменшити надто великі значення бета, оскільки високі значення бета збільшують значення функції втрат залежно від коливань лямбда.

$$\underbrace{\hat{\beta} \equiv \underset{\beta}{argmin}(\|y - X\beta\|^2)}_{Original\ min.problem} \Rightarrow \underbrace{\hat{\beta} \equiv \underset{\beta}{argmin}(\|y - X\beta\|^2 + \lambda_1\|\beta\| + \lambda_2\|\beta\|^2)}_{Elastic\ Net\ min.problem}$$

Регресія LASSO, LASSO Regression – окремий випадок Еластичної сітки, де $\lambda_1 = \lambda$, $\lambda_2 = 0$.

Регресія з регуляризацією Тихонова, Ridge regression – окремий випадок Еластичної сітки, де $\lambda_1 = 0$, $\lambda_2 = \lambda$.

Бегінг, Bagging – техніка комбінації, відповідно до якої загальний набір даних рівномірно розподіляється на підвибірки, а моделі тренуються на кожній підвибірці. Тоді отримані коефіцієнти необхідно об'єднати за допомогою певного усереднення (залежно від конкретного методу та задачі, до яких застосовується бегінг).

Дерева рішень, Decision Trees – алгоритм побудови дерева (графіка), де кожен нейрон становить "запитання" до характеристик спостереження. Відповіді на ці запитання ведуть до листків, які становлять певне значення чи клас.

Випадковий ліс, Random Forest (RF) – поєднання техніки бегінгу та дерева рішень, де дерева будуються для різних підвибірок характеристик, а пізніше об'єднуються.

G Метод градієнтного бустингу, radient Boosting Method (GBM) – ансамблевий алгоритм, відповідно до якого стверджується, що об'єднавши результат кількох слабких підмоделей, можна отримати гарне рішення. Він тренує слабкі моделі та ітераційно додає їх до сильної комбінації. Під час кожної ітерації дані повторно зважуються і надають більшої ваги тим даним, прогноз щодо яких раніше був гірший.

EXtreme Gradient Boosting (XGBoost) – бібліотека з відкритим кодом, яка пропонує систему градієнтного підсилювання. Вона стала дуже популярною та має переваги порівняно з іншими бібліотеками (LightGBM, CatBoost). Однак вона також має деякі недоліки, залишаючи відкритим питання, яка бібліотека все-таки краща.

Super Learner – алгоритм, заснований на методі стекінгу, який є третьою основною технікою ансамблю методів. Він складається з двох етапів: навчання великої кількості слабких підмоделей (необов'язково однотипних, тобто це можуть бути різні техніки); навчання мета-моделі, яка використовує результати цих моделей для реального прогнозування.

Кластеризація, Clustering – сукупність інструментів для групування об'єктів за їх подібністю.

Метод k-середніх, K-Means – один із найпопулярніших алгоритмів кластеризації. Він довільно присвоює k точкам значення центрів кластера, а потім ітеративно переміщує їх до центру початкової та найближчих точок. Цей підхід мінімізує дисперсію всередині кластера.

Метод опорних векторів, Support Vector Machine (SVM) – модель для побудови гіперплощини для розділення спостережень на кілька груп. Він максимізує відстань від гіперплощини до найближчих спостережень з обох сторін (встановлює розділення).

k-найближчі сусіди, k-Nearest Neighbours (k-NN) – модель класифікації. Відповідно до цієї моделі новому об'єкту присвоюється клас залежно від того, скільки об'єктів цього класу загалом знаходиться в k найближчих точках.

Наївний Байєс, Naïve Bayes – це теорема Байєса, яка використовується для даних із припущенням, ознаки якого роблять незалежний внесок до ймовірності події (тому її називають наївною).

Зниження розмірності, Dimensionality reduction – набір інструментів для зменшення розмірності простору (кількості ознак) без значної втрати інформації.

Нейронна мережа, Neural Network (NN) – цей алгоритм складається з кількох частин. Є нейрони, шари, з'єднання та функції активації. Нейрони складаються з деяких значень, і вони утворюють упорядковані шари. Усі нейрони першого шару з'єднані з усіма нейронами другого, нейрони другого шару – з нейронами третього і т. д. Перший шар має вхідні дані, а останній дає вихідні дані. Значення нейрона дорівнює зваженій сумі значень у всіх нейронах попереднього шару (під'єднаних нейронів), перетворених за допомогою функції активації (яка традиційно виробляє значення від 0 до 1, як у логістичній регресії). Метою алгоритму є калібрування ваг таким чином, щоб мінімізувати відхилення між установленою та реальною продуктивністю.

Глибока нейронна мережа, Deep Neural Network (DNN) – це нейронна мережа, але з набагато більшою кількістю шарів. У деяких випадках це значно підвищує точність, але підходи до роботи з алгоритмом теж дещо відрізняються.

Рекурентна нейронна мережа, Recurrent Neural Network (RNN) – це нейронна мережа, де деякі нейрони можуть повторюватися. Це означає, що вона приймає значення не лише попереднього шару, а й власне значення (що відповідає визначенню рекурентності).

Латентний розподіл Дірихле, Latent Dirichlet Allocation (LDA) – алгоритм, що застосовується здебільшого для обробки природної мови та групування текстів за темами. Відповідно до цього алгоритму набір тем розподіляється за набором документів та набором слів для зменшення кількості прямих зв'язків згідно з розподілом Діріхле.

Двонаправлений кодувальник на основі штучного інтелекту, Bidirectional Encoder Representation from Transformers (BERT) – провідна техніка для різних завдань NLP (на момент публікації). Він використовує трансформатори, які зчитують усю послідовність слів, вставляють їх у вектори, замінюють деякі слова маскувальним токеном (для поліпшення контекстного навчання) і передають їх у нейронну мережу.