

РОЗВІДУВАЛЬНИЙ АНАЛІЗ ДАНИХ ДЛЯ ІНФОРМАЦІЙНОЇ ТЕХНОЛОГІЇ ПЕРЕДБАЧЕННЯ ЗВІЛЬНЕННЯ СПІВРОБІТНИКІВ В ОРГАНІЗАЦІЯХ МЕТОДАМИ МАШИННОГО НАВЧАННЯ

Вінницький національний технічний університет

Анотація

Робота присвячена підготовці та розвідувальному аналізу даних для подальшого використання для інформаційної технології передбачення звільнення співробітників в організаціях методами машинного навчання. Було проведено аналіз датасету та його ознак.

Ключові слова: звільнення, інформаційні технології, машинне навчання, аналіз даних, передбачення, ознаки, звільнення співробітників.

Abstract

The paper is devoted to the preparation and exploratory analysis of data for further use in the information technology of predicting the dismissal of employees in organizations using machine learning methods. The dataset and its features were analyzed.

Keywords: dismissal, information technology, machine learning, data analysis, predictions, signs, employee dismissal.

Вступ

Плинність кадрів є комплексним феноменом, який має як кількісні, так і якісні виміри. Вона може бути природною та навіть корисною, наприклад, коли організацію залишають малоефективні співробітники, або ж коли відбувається приплив свіжих ідей та компетенцій. Проте, непередбачуване та масове звільнення цінних фахівців, відоме як "вигорання" або "відтік талантів", становить серйозну загрозу для стабільності та конкурентоспроможності компанії.

Особливо гостро проблема плинності кадрів стоїть в ІТ-галузі, де спостерігається високий попит на кваліфікованих фахівців, і конкуренція за них є дуже високою. "Переманювання" спеціалістів, швидкі зміни технологій, прагнення до саморозвитку та постійного кар'єрного зростання роблять ІТ-співробітників більш схильними до зміни місця роботи. Це підкреслює необхідність для ІТ-компаній мати ефективні інструменти для передбачення та управління плинністю кадрів.

Таким чином, у зв'язку з вищезазначеними наслідками та викликами, виникає нагальна потреба в розробці інструментів, які дозволять не лише реєструвати факт звільнення, але й прогнозувати його з високою ймовірністю. Інформаційна технологія, що базується на методах машинного навчання, здатна аналізувати великі обсяги даних про співробітників (демографічні дані, історію роботи, продуктивність, участь у проєктах, опитування, відгуки тощо) та виявляти приховані закономірності, які передують звільненню. Це дозволить організаціям перейти від реактивного до проактивного управління персоналом, вчасно втручатися та вживати заходів для утримання цінних співробітників, тим самим мінімізуючи негативні наслідки плинності кадрів та підвищуючи загальну ефективність бізнесу [1,2].

Розвідувальний аналіз

Для проведення аналізу було обрано набір даних у Kaggle, що має назву «Employee dataset» та опублікований користувачем «Tawfik Elmetwally». Даний датасет має відкритий доступ для загального використання на платформі [3]. Цей набір даних містить інформацію про співробітників компанії, включаючи їхню освіту, трудову діяльність, демографічні дані та фактори, пов'язані з працевлаштуванням. Він анонімізований для захисту конфіденційності,

водночас надаючи цінну інформацію про робочу силу. Всього в датасеті 9 ознак та 4653 записів (рис.1).

	Education	JoiningYear	City	PaymentTier	Age	Gender	EverBenched	ExperienceInCurrentDomain	LeaveOrNot
0	Bachelors	2017	Bangalore	3	34	Male	No	0	0
1	Bachelors	2013	Pune	1	28	Female	No	3	1
2	Bachelors	2014	New Delhi	3	38	Female	No	2	0
3	Masters	2016	Bangalore	3	27	Male	No	5	1
4	Masters	2017	Pune	3	24	Male	Yes	2	1
...	...	...	...	...	...	...	...	...	...
4648	Bachelors	2013	Bangalore	3	26	Female	No	4	0
4649	Masters	2013	Pune	2	37	Male	No	2	1
4650	Masters	2018	New Delhi	3	27	Male	No	5	1
4651	Bachelors	2012	Bangalore	3	30	Male	Yes	2	0
4652	Bachelors	2015	Bangalore	3	33	Male	Yes	4	0

Рис. 1. Приклад ознак у наборі даних

Даний датасет включає у себе такі колонки:

1. Education: Освітня кваліфікація працівників, включаючи ступінь, навчальний заклад та галузь навчання.
2. Joining Year: Рік, коли кожен працівник приєднався до компанії, із зазначенням стажу роботи.
3. City: Місцезнаходження або місто, де базується або працює кожен працівник.
4. Payment Tier: Класифікація працівників за різними рівнями оплати праці.
5. Age: Вік кожного працівника, що надає демографічну інформацію.
6. Gender: Гендерна ідентичність працівників, що сприяє аналізу різноманітності.
7. Ever Benched: Вказує, чи був працівник коли-небудь тимчасово без призначеної роботи.
8. Experience in Current Domain: Кількість років досвіду роботи працівників у їхній поточній галузі.
9. Leave or Not: цільовий стовпець

Використано функцію describe(), щоб визначити описові статистики числових ознак (рис.2).

Описова статистика числових ознак:					
	JoiningYear	PaymentTier	Age	ExperienceInCurrentDomain	LeaveOrNot
count	4653.000000	4653.000000	4653.000000	4653.000000	4653.000000
mean	2015.062970	2.698259	29.393295	2.905652	0.343864
std	1.863377	0.561435	4.826087	1.558240	0.475047
min	2012.000000	1.000000	22.000000	0.000000	0.000000
25%	2013.000000	3.000000	26.000000	2.000000	0.000000
50%	2015.000000	3.000000	28.000000	3.000000	0.000000
75%	2017.000000	3.000000	32.000000	4.000000	1.000000
max	2018.000000	3.000000	41.000000	7.000000	1.000000

Рис. 2. Результат виконання функції describe()

З рисунка 2 видно, що існує помірний дисбаланс класів, оскільки клас "0" (близько 65.61%) зустрічається частіше, ніж клас "1" (34.39%). Це необхідно враховувати при виборі метрик оцінки моделі (наприклад, F1-міра, Precision, Recall замість лише точності) та, можливо, застосувати методи боротьби з дисбалансом класів (наприклад, оверсемплінг, андерсемплінг). Візуалізуємо розподіл цільової змінної (рис.3).

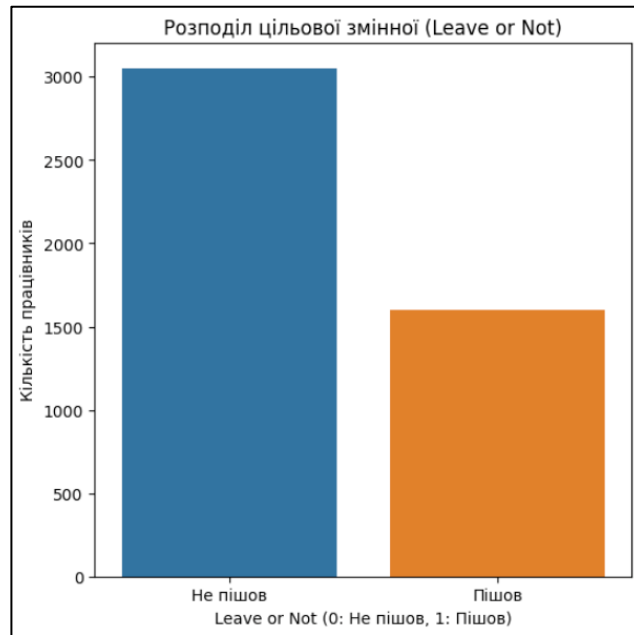


Рис. 3. Розподіл цільової змінної

З рисунка 3 видно помірний дисбаланс класів – кількість працівників, які не пішли, майже вдвічі більша за кількість тих, хто пішов.

Візуалізуємо тривалість роботи (Joining Year) в різних містах (рис.4).

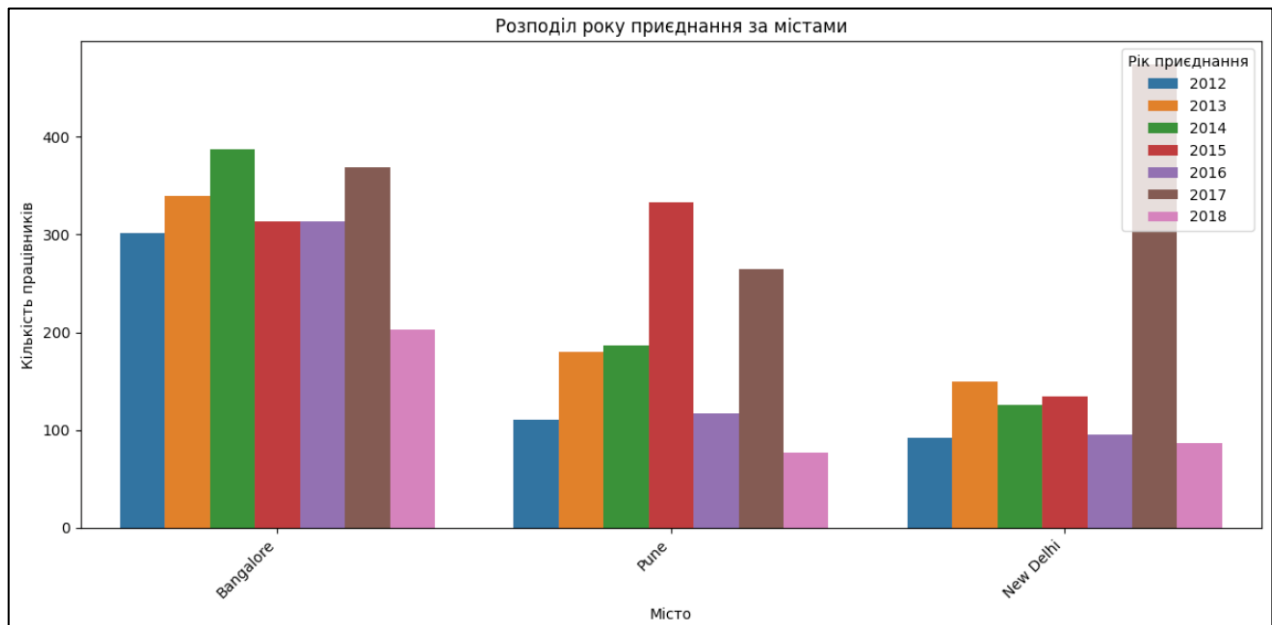


Рис. 4. Розподіл освітніх кваліфікацій серед працівників

З рисунка 4 видно, що Bangalore є лідером за кількістю нових працівників у більшості років, особливо помітні піки у 2014, 2016 та 2017 роках. У 2018 році у всіх містах спостерігається значне зниження кількості приєднань порівняно з попередніми роками, 2016 та 2017 роки були досить активними для найму, особливо у Bangalore та New Delhi.

Далі розглянемо зв'язок цільової змінної з освітою (рис.5).

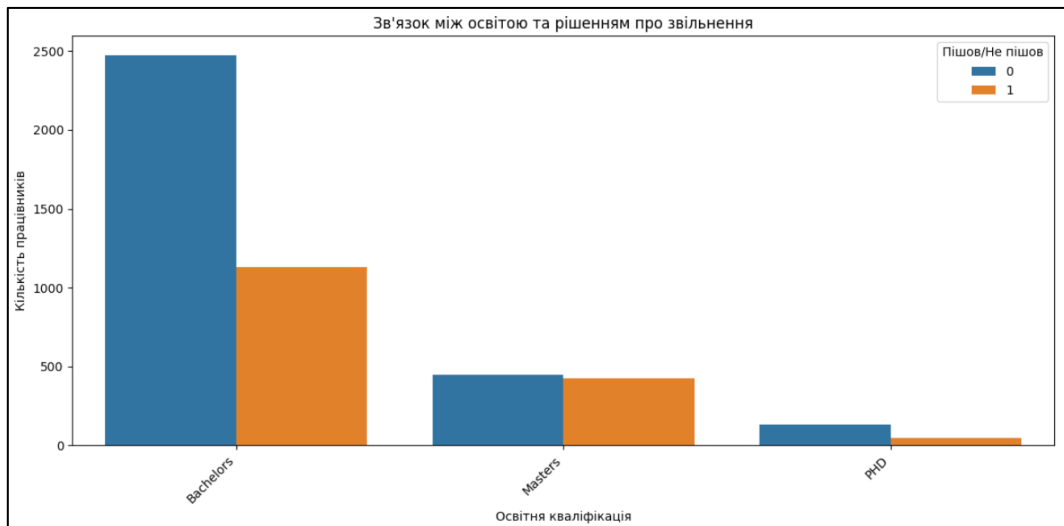


Рис. 5. Зв'язок між освітою та рішенням про звільнення

Як можна побачити, існує чіткий зв'язок між рівнем освіти та рішенням про звільнення. Працівники з нижчою освітою (Bachelors) становлять найбільшу групу тих, хто звільняється в абсолютних цифрах, тоді як магістри мають найвищий відсоток звільнень відносно своєї групи. На противагу цьому, працівники з PhD є найбільш стабільними з точки зору звільнень.

Розглянемо зв'язок цільової змінної з рівнем оплати (рис.6).

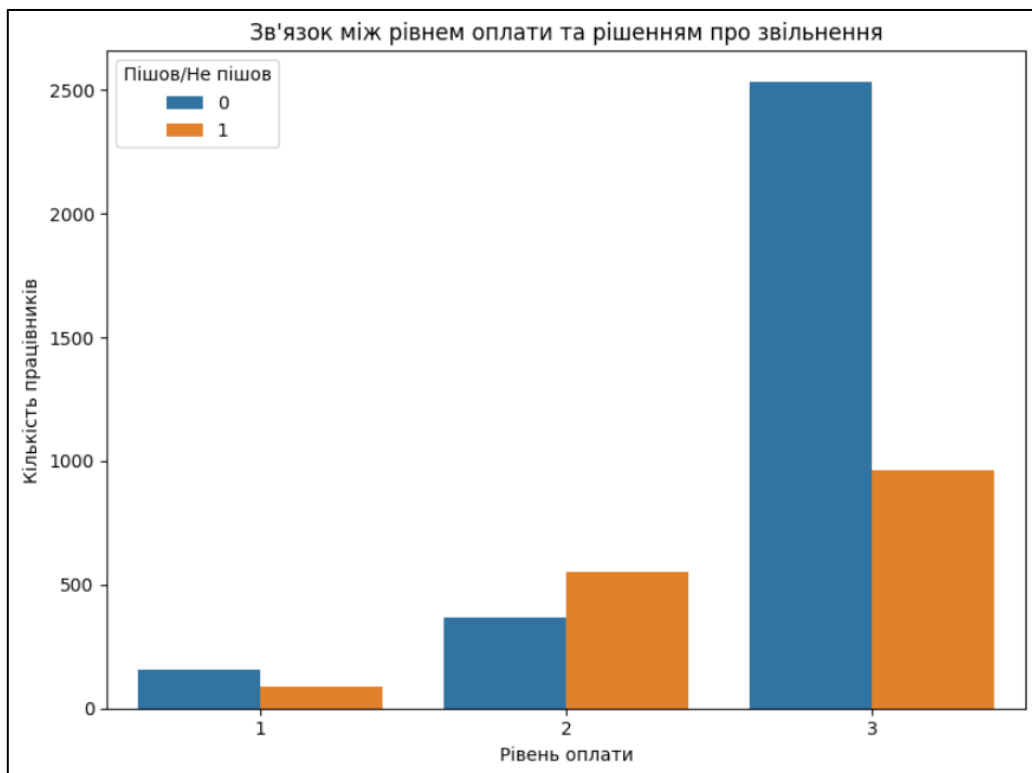


Рис. 6. Зв'язок між рівнем оплати та рішенням про звільнення

Цей графік чітко демонструє, що рівень оплати праці є значущим фактором, що впливає на рішення співробітників щодо звільнення.

Чим вищий рівень оплати, тим менша відносна частка звільнень. Це підкреслює важливість конкурентної заробітної плати для утримання кваліфікованих кадрів.

Незважаючи на це, звільнення відбуваються на всіх рівнях оплати, навіть на найвищому. Це означає, що, окрім фінансового стимулу, існують й інші нефінансові фактори (такі як задоволеність роботою, кар'єрний ріст, культура компанії, відносини з керівництвом, баланс робота/життя тощо), які також впливають на рішення про відхід.

На рисунку 7 продемонстровано матрицю кореляції та їх теплова мапа.

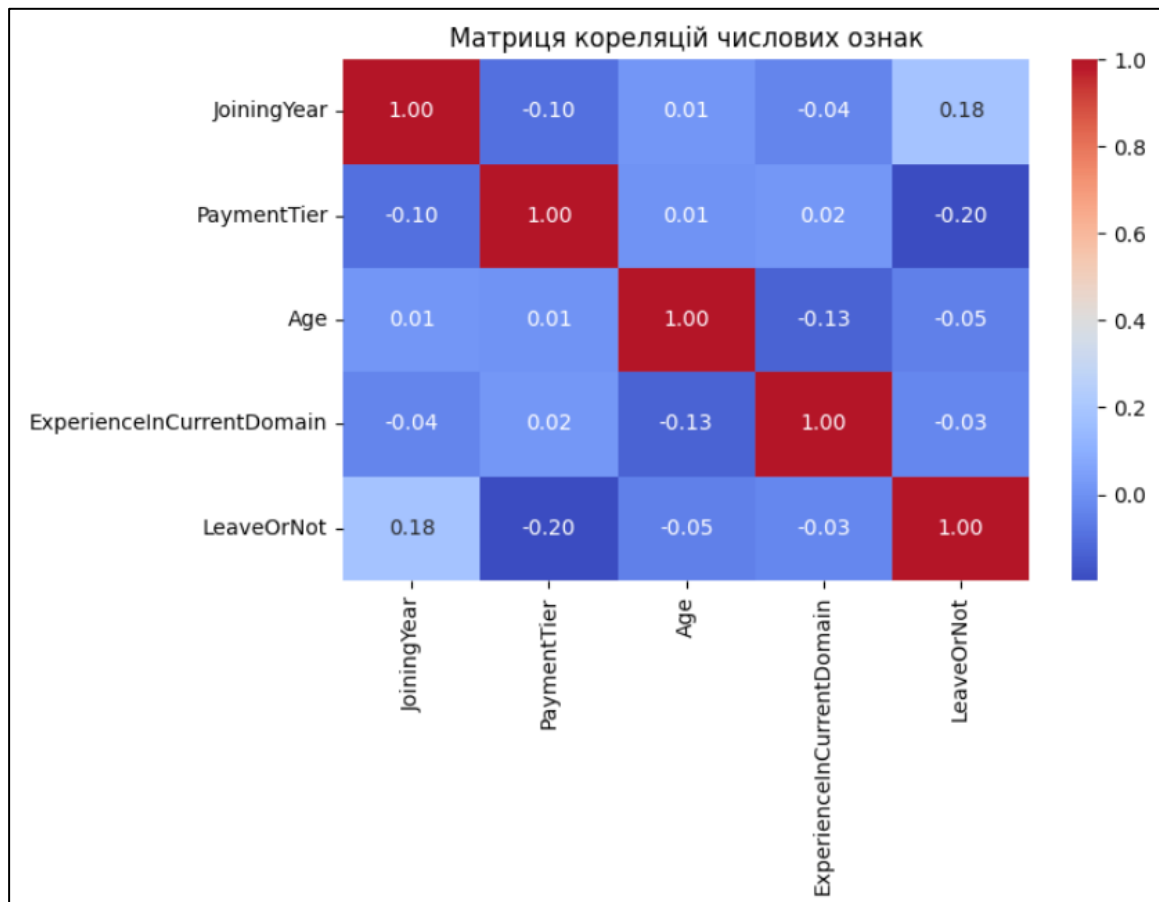


Рис. 7. Теплова мапа матриці кореляції

З рисунку 7 можна зробити висновок що PaymentTier є найсильнішим числовим предиктором звільнення серед представлених числових ознак, хоча сила кореляції все ще є помірною. Це підтверджує гіпотезу, що фінансовий фактор є важливим, але не єдиним чинником.

JoiningYear також має слабкий, але помітний зв'язок зі звільненням, що може вказувати на певні тенденції щодо відтоку нових співробітників.

Age та ExperienceInCurrentDomain демонструють дуже слабкі лінійні кореляції зі звільненням. Це не означає, що вони зовсім не важливі, але їхній лінійний вплив на рішення про звільнення є мінімальним у даному наборі даних. Можливо, їх вплив є нелінійним або проявляється через взаємодію з іншими змінними.

Відсутність сильної мультиколінеарності між незалежними числовими ознаками є позитивним фактором для побудови моделей машинного навчання.

Висновки

Проведений розвідувальний аналіз даних дозволив отримати глибоке розуміння структури та характеристик набору даних, що використовуються для прогнозування кількості продажів.

Ідентифіковано аномальні значення, які потребуватимуть подальшої обробки для запобігання викривленню результатів моделювання.

Виявлено нерівномірний розподіл даних між категоріями, що може вплинути на продуктивність моделі і потребуватиме застосування методів балансування.

Встановлено взаємозв'язки між різними параметрами, що є ключовим для вибору релевантних ознак та побудови ефективної моделі прогнозування.

Визначено, що основними факторами, які чинять найбільший вплив на кількість продажів, є роздрібна ціна, місяць року, розмір знижки та сегмент, у якому відбувається продаж. Це підтверджує, що економічні та часові показники є вирішальними для формування прогнозів.

Результати EDA стануть основою для подальшої передобробки даних та вибору оптимальних алгоритмів машинного навчання, спрямованих на максимізацію точності передбачення звільнення співробітників.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Raza, A.; Munir, K.; Almutairi, M.; Younas, F.; Fareed, M.M.S. Predicting Employee Attrition Using Machine Learning Approaches. *Appl. Sci.* 2022, 12, 6424. <https://doi.org/10.3390/app12136424>
2. Norsuhada Mansor, Nor Samsiah Sani and Mohd Aliff, "Machine Learning for Predicting Employee Attrition" *International Journal of Advanced Computer Science and Applications(IJACSA)*, 12(11), 2021. <http://dx.doi.org/10.14569/IJACSA.2021.0121149>
3. Employee Dataset. Kaggle. URL: <https://www.kaggle.com/datasets/tawfikelmetwally/employee-dataset/data>
4. Matplotlib Pyplot Documentation. 2025. URL: https://matplotlib.org/3.5.3/api/as_gen/matplotlib.pyplot.html
5. Seaborn Tutorial. 2025. URL: <https://seaborn.pydata.org/tutorial.html>

Кордонський Арсен Олександрович – студент групи 2ICT-216, факультет інтелектуальних інформаційних технологій та автоматизації, Вінницький національний технічний університет, Вінниця, e-mail: arsen.14.sloboda@gmail.com

Козачко Олексій Миколайович – к.т.н., доцент кафедри системного аналізу та інформаційних технологій, Вінницький національний технічний університет, e-mail: lekoz80@gmail.com

Kordonsky Arsen O. - student of Faculty of Intelligent Information Technology and Automation, 2IST-21b, Vinnytsia National Technical University, Vinnytsia, e-mail arsen.14.sloboda@gmail.com

Kozachko Oleksiy M. - Ph.D., Assistant Professor of the Department of System Analysis and Information Technologies, Vinnytsia National Technical University, Vinnytsia, e-mail: lekoz80@gmail.com