МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний університет «Острозька академія»
Навчально-науковий інститут інформаційних технологій та бізнесу
Кафедра інформаційних технологій та аналітики даних

КВАЛІФІКАЦІЙНА РОБОТА

на здобуття освітнього ступеня магістра

на тему: **«Управління проєктом з розробки моделі утримання працівників ІТ компанії»**

Виконав: студент 2 курсу, групи МУП-21
другого (магістерського) рівня вищої освіти
спеціальності 122 Комп'ютерні науки
освітньо-професійної програми «Управління проєктами»
Ягодка Андрій Вікторович

Керівник: старший викладач
Клебан Юрій Вікторович

Рецензент: кандидат технічних наук, доцент, доцент кафедри прикладної математики Донецького національного університету імені Василя Стуса Загоруйко Любов Василівна

РОБОТА ДОПУЩЕНА ДО ЗАХИСТУ
Завідувач кафедри інформаційних технологій та аналітики даних
_____ (проф., д.е.н. Кривицька О.Р.)
Протокол № 5 від 04 грудня 2025 р.

Острог, 2025

АНОТАЦІЯ

кваліфікаційної роботи
на здобуття освітнього ступеня магістра

Тема: Управління проєктом з розробки моделі утримання працівників IT компанії

Автор: Ягодка Андрій Вікторович

Науковий керівник: старший викладач Клебан Юрій Вікторович

Захищена «……»…… 2025 року.

Пояснювальна записка до кваліфікаційної роботи: 126 с., 38 рис., 0 табл., 1 додаток, 48 джерел.

Ключові слова: управління проєктами, утримання працівників, IT-компанія, машинне навчання, HR-аналітика, прогнозування звільнень

Короткий зміст праці:

Магістерська кваліфікаційна робота присвячена управлінню проєктом із розробки аналітичної моделі для прогнозування утримання працівників IT-компанії. У роботі розглянуто сучасні підходи до управління проєктами у сфері Data Science, специфіку утримання працівників у IT-галузі та методи машинного навчання, що можуть бути використані для передбачення звільнень.

На основі аналізу наукових джерел і практик HR-аналітики сформовано методологію розробки моделі прогнозування відтоку працівників. У дослідженні використано алгоритми машинного навчання, зокрема логістичну регресію, Random Forest та XGBoost, для визначення факторів, які найбільше впливають на лояльність персоналу. Проведено оцінку точності моделей та розраховано економічний ефект від їх впровадження.

Робота містить етапи управління проєктом – від ініціації до впровадження аналітичної системи у процеси HR-менеджменту. Запропоновано рекомендації щодо інтеграції моделі в діяльність IT-компаній для підвищення ефективності управління персоналом та зменшення плинності кадрів. Результати дослідження можуть бути використані HR-відділами для переходу від реактивного до проактивного управління персоналом, що сприятиме підвищенню стабільності та конкурентоспроможності організації.

ABSTRACT
of the qualification work
for obtaining a master's degree

Topic: Project Management for the Development of an Employee Retention Model in an IT Company

Author: Andrii Yahodka

Scientific supervisor: Senior Lecturer Yurii Kleban

Protected «……»…… 2025 .

Explanatory note to the qualification work: *126 p., 38 figures, 0 tables, 1 appendic, 48 sources.*

Keywords: project management, employee retention, IT company, machine learning, HR analytics, employee turnover prediction

Summary of the work:

The master's qualification work is devoted to project management in the development of an analytical model for predicting employee retention in an IT company. The research explores modern approaches to project management in the Data Science domain, specific features of employee retention in the IT industry, and the application of machine learning methods for employee turnover prediction.

Based on the analysis of scientific literature and HR analytics practices, a methodology for developing a predictive model of employee attrition was proposed. Machine learning algorithms such as Logistic Regression, Random Forest, and XGBoost were applied to identify key factors influencing employee loyalty. The models' accuracy and the potential economic effect of their implementation were evaluated.

The work presents all stages of project management. From initiation to the implementation of the analytical system within HR processes. Recommendations were developed for integrating the model into IT company operations to improve personnel management efficiency and reduce employee turnover. The obtained results can be used by HR departments to transition from reactive to proactive workforce management, thereby enhancing organizational stability and competitiveness.

ЗМІСТ

ВСТУП

Сфера інформаційних технологій (IT) є однією з найбільш динамічних та конкурентних галузей сучасної економіки. Високий рівень конкуренції за талановитих фахівців, швидкий розвиток технологій та зміна пріоритетів працівників зумовлюють необхідність розробки ефективних стратегій утримання персоналу. Втрата кваліфікованих спеціалістів не лише створює ризики для стабільності та продуктивності компанії, а й т ягне за собою значні фінансові витрати, пов'язані з пошуком, адаптацією та навчанням нових співробітників [1].

Згідно з дослідженнями галузевих експертів, вартість заміщення одного IT-спеціаліста може становити від 50% річної заробітної плати, залежно від рівня кваліфікації та посади. Особливо критичною є ситуація з розробниками високого рівня та архітекторами систем, де період пошуку та адаптації нового фахівця може тривати більш ніж 3 місяці, що призводить до затримок у виконанні проєктів [5].

У сучасній IT-індустрії утримання кваліфікованих працівників є критично важливим завданням для забезпечення стабільності та конкурентоспроможності компаній. Висока плинність кадрів призводить до значних фінансових витрат, пов'язаних із пошуком, наймом та навчанням нових співробітників, а також може негативно впливати на моральний стан колективу та продуктивність роботи. Статистичні дані показують, що середній рівень плинності кадрів в IT-секторі становить 18-25% на рік, що перевищує показники інших галузей економіки 10-15%.

Сучасні HR-фахівці стикаються з низкою викликів, пов'язаних з управлінням персоналом в умовах цифрової трансформації. Зростання популярності віддаленої роботи, поширення гібридних моделей зайнятості та підвищені очікування працівників щодо балансу життя та роботи створюють нові виклики для традиційних підходів до утримання. Крім того, швидка еволюція технологічного стеку вимагає постійного професійного розвитку співробітників, що також впливає на їхню лояльність до компанії.

Традиційні методи управління персоналом часто не дозволяють своєчасно виявляти працівників, схильних до звільнення, або розуміти глибинні причини

їхнього незадоволення. У зв'язку з цим, застосування методів машинного навчання для прогнозування звільнень стає все більш актуальним. Наприклад, у дослідженні Діно Майкла Квінтероса було розроблено оптимізовану модель прогнозування відтоку працівників за допомогою глибокого навчання, яка досягла точності прогнозування [2].

Дослідження К'яра Мореллі, Джанлука Фузай та Раффаеле Зенті [3] підкреслюють важливість інтеграції методів машинного навчання в управління персоналом. Автори наголошують на необхідності ретельної обробки даних, вибору релевантних характеристик та використання ансамблевих моделей для підвищення точності прогнозів. Вони також ідентифікували десять ключових факторів, що впливають на плинність кадрів, що дозволяє створити профіль працівника з високою ймовірністю звільнення. Це дослідження демонструє потенціал машинного навчання у формуванні ефективних HR-стратегій, спрямованих на зниження плинності кадрів та підвищення задоволеності працівників.

Аналіз міжнародного досвіду показує, що компанії, які впроваджують машинне навчання в HR-аналітику, демонструють на 15-30% нижчий рівень плинності кадрів порівняно з організаціями, що використовують лише традиційні методи управління персоналом [23]. Водночас, успішність таких ініціатив значною мірою залежить від якості даних, правильності вибору алгоритмів та ефективності управління проєктом їх розробки та впровадження.

Дослідження Сувой Редді та Лідії Дівія [4] порівнює ефективність різних алгоритмів побудови моделей, таких як Naive Bayes, Random Forest та XGBoost, у прогнозуванні звільнень працівників. Результати показали, що модель XGBoost досягла найвищої точності прогнозування 85,3%, що підкреслює потенціал використання передових алгоритмів для підвищення точності прогнозів у сфері управління персоналом.

Важливим аспектом, який часто залишається поза увагою дослідників, є психологічні та соціальні чинники, що впливають на рішення працівника залишити компанію. У дослідженні під назвою «Вплив звільнення колег на наміри щодо плинності кадрів» [5] розглядається вплив звільнень колег на наміри інших

працівників залишити компанію. Автори виявили, що звільнення колег значно підвищує сприйняття зовнішніх можливостей працевлаштування та збільшує наміри щодо звільнення серед інших співробітників. Це підкреслює важливість врахування соціальних взаємодій при розробці моделей прогнозування плинності кадрів.

Особливої уваги заслуговує специфіка IT-галузі, де працівники часто мають унікальні навички та можуть легко змінювати роботодавців. Дослідження показують, що основними факторами незадоволення IT-спеціалістів є: відсутність можливостей для професійного зростання, технологічна відсталість проєктів, неконкурентна заробітна плата, конфлікти в команді та недостатня автономія у прийнятті технічних рішень [17].

Таким чином, використання моделей машинного навчання для прогнозування звільнень є актуальним напрямом досліджень, який має значний потенціал для підвищення ефективності управління персоналом у компаніях. Водночас, розробка та впровадження таких моделей вимагає системного підходу до управління проєктом, що включає планування, організацію ресурсів, контроль якості та ризик-менеджмент.

Метою мого дослідження є розробка та впровадження моделі утримання працівників IT-компанії на основі методів машинного навчання з використанням принципів проєктного менеджменту, що дозволить виявити ключові фактори, які впливають на рівень задоволеності та лояльності персоналу, а також створити ефективну систему раннього попередження про ризики звільнення співробітників.

Загалом побудова моделі утримання працівників потребуватиме виконання наступних завдань:

- проаналізувати теоретичні аспекти утримання персоналу в IT-компаніях та сучасні підходи до управління проєктами;
- визначити основні фактори, що впливають на рішення працівників залишити компанію, з урахуванням специфіки IT-галузі;
- сформувати комплексний підхід до управління проєктом розробки аналітичної моделі утримання працівників, включаючи планування ресурсів та управління ризиками;

- зібрати та підготувати дані для побудови прогнозної моделі з дотриманням принципів захисту персональних даних;
- розробити та оцінити декілька моделей машинного навчання для передбачення ризиків звільнення з порівняльним аналізом їх ефективності;
- розробити детальні рекомендації щодо впровадження моделі в HR-процеси компанії та оцінити економічну ефективність проєкту;
- створити план моніторингу та підтримки моделі для забезпечення її довгострокової ефективності.

Об'єктом дослідження є процес управління проєктом з розробки аналітичної моделі утримання працівників IT-компанії, включаючи всі етапи від ініціації до закриття проєкту та впровадження моделі у використання.

Предметом дослідження виступають методи управління проєктами в сфері роботи з даними, аналітичні інструменти для прогнозування звільнень працівників, а також організаційні процеси інтеграції машинного навчання в HR-діяльність IT-компаній.

Практичне значення дослідження полягає у можливості використання розробленої моделі для підвищення рівня утримання працівників у IT-компаніях. Запропонований підхід може бути адаптований до потреб різних організацій, що дозволить HR-фахівцям приймати обґрунтовані рішення щодо управління персоналом та зменшити ризики втрати ключових співробітників.

Результати дослідження матимуть практичну цінність на різних рівнях. Для HR-департаментів IT-компаній розроблена модель надасть можливість переходу від реактивного до проактивного управління персоналом, дозволяючи виявляти потенційні проблеми на ранніх стадіях та розробляти персоналізовані стратегії утримання. Керівництво компаній зможе приймати більш обґрунтовані рішення щодо інвестицій у розвиток персоналу та оптимізації HR-бюджету.

Для проєктних менеджерів результати дослідження представлятимуть цінність у контексті розуміння специфіки управління проєктами та інтеграції аналітичних рішень у бізнес-процеси. Розроблена методологія може служити основою для

створення стандартизованих процедур управління аналогічними проєктами в майбутньому.

Економічний ефект від впровадження моделі може бути суттєвим. За оцінками експертів, зниження плинності кадрів навіть на 5-10% може призвести до економії коштів у розмірі 15-25% від річного HR-бюджету компанії за рахунок зменшення витрат на рекрутинг, онбординг та втрачену продуктивність [23].

Крім безпосередніх економічних переваг, впровадження предиктивної моделі сприятиме покращенню організаційної культури та підвищенню рівня довіри між співробітниками та менеджментом. Використання даних для прийняття рішень демонструє прозорість та об'єктивність HR-процесів, що є особливо важливим для IT-спеціалістів, які цінують логічний та аналітичний підхід до вирішення проблем.

Таким чином, використання аналітичних підходів та методів машинного навчання в управлінні персоналом IT-компаній дозволяє не лише передбачати ймовірність звільнень, а й формувати ефективні стратегії утримання працівників. Реалізація таких моделей сприятиме зниженню плинності кадрів, покращенню рівня задоволеності працівників та забезпеченню стабільності бізнес-процесів.

Подальші дослідження можуть бути спрямовані на вдосконалення прогнозних моделей, зокрема шляхом розширення набору змінних, що враховують соціальні, психологічні та професійні аспекти діяльності співробітників. Перспективним напрямом є інтеграція моделей прогнозування з системами автоматизованого HR-менеджменту та розробка інтерактивних графіків для візуалізації результатів аналізу.

Крім того, перспективним напрямом є розробка інтегрованих аналітичних систем з управління персоналом, які автоматично генеруватимуть рекомендації для HR-менеджерів щодо персоналізованих заходів утримання працівників. Такі системи можуть включати модулі для аналізу настроїв співробітників на основі текстових даних, моніторингу продуктивності та прогнозування кар'єрних траєкторій.

Запропонована методика може бути адаптована для різних сегментів ринку праці, що дозволить підвищити ефективність управління персоналом у широкому спектрі компаній, а також сприятиме подальшому розвитку науки управління людськими ресурсами на основі сучасних аналітичних технологій. Особливу увагу

слід приділити етичним аспектам використання персональних даних працівників та забезпеченню прозорості алгоритмів прийняття рішень.

РОЗДІЛ 1. ТЕОРЕТИЧНІ ЗАСАДИ УПРАВЛІННЯ УТРИМАННЯМ ПРАЦІВНИКІВ ІТ КОМПАНІЇ

1.1. Сучасні підходи до управління проєктами в сфері Data Science

У сучасному бізнес-середовищі управління проєктами в сфері машинного навчання та роботи з даними особливого значення через специфіку роботи з даними та необхідність забезпечення високої якості аналітичних рішень. Розробка моделі утримання працівників ІТ компанії як проєкт вимагає застосування сучасних методологій управління проєктами, адаптованих до особливостей роботи з великими обсягами даних та машинним навчанням. Традиційні підходи до управління проєктами, що були розроблені для класичних ІТ-систем або інженерних рішень, часто виявляються недостатньо ефективними в контексті аналітичних проєктів, де результат значною мірою залежить від якості даних та непередбачуваності процесу дослідження.

Для розуміння еволюції управління проєктами у сфері роботи з даними та розробки моделей машинного навчання доцільно проаналізувати традиційні підходи, зокрема Waterfall, а також сучасні гнучкі методології, об’єднані під спільною назвою Agile.

Waterfall, або каскадна модель, є класичною лінійною методологією управління проєктами, яка передбачає проходження через чітко визначені етапи: визначення вимог, проєктування, розробка, тестування, впровадження та супровід. Кожна фаза має бути завершена повністю до початку наступної, що унеможливлює гнучке повернення до попередніх етапів без істотних часових та ресурсних втрат. Концепція каскадної моделі була вперше запропонована Вінстоном Ройсом у 1970-х роках, однак уже тоді він сам відзначав її обмеження, зокрема відкладене тестування, що часто призводить до виявлення критичних помилок надто пізно в життєвому циклі проєкту [6].

Серед переваг Waterfall варто відзначити чіткість структури, передбачуваність термінів і бюджету, детальне планування та документацію, що робить її ефективною для проєктів зі стабільними, незмінними вимогами. Такий підхід підходить для

регламентованих сфер, наприклад, у будівництві чи проєктуванні апаратного забезпечення, де зміни в процесі реалізації мають бути мінімізовані. Проте у сфері роботи з даними та машинним навчання, яка передбачає високу ступінь невизначеності, експериментальний характер роботи та часті зміни вимог залежно від результатів аналізу даних, Waterfall виявляється малоефективною. Неможливість швидко реагувати на нові дані, результати експериментів або зміну бізнес-цілей суттєво знижує гнучкість цього підходу в таких проєктах [6].

У відповідь на обмеження традиційних підходів у 2001 році було сформульовано новий напрям Agile підходів, який започаткував нову еру в управлінні проєктами. Agile об’єднує низку методологій (Scrum, Kanban та інші), що базуються на спільних принципах: фокус на цінність для замовника, гнучкість до змін, самоуправління команд, ітеративність розробки та регулярне постачання працюючого продукту. Згідно з цими принципами, ключові цінності Agile включають [7]:

- люди та взаємодії понад процеси та інструменти;
- працююче програмне забезпечення важливіше за ідеально структуровану документацію;
- співпраця із замовником понад контрактні домовленості;
- готовність до змін понад дотримання первинного плану.

Agile передбачає реалізацію проєкту через низку коротких ітерацій (спринтів), тривалістю зазвичай від 1 до 4 тижнів. Кожен спринт завершується створенням або оновленням функціонального компонента продукту, що дозволяє швидко перевірити гіпотези, отримати зворотний зв’язок від користувача чи замовника та оперативно ввести корективи. Такий підхід забезпечує гнучкість, адаптивність до нових обставин і можливість швидкого масштабування рішень у відповідь на зміну бізнес-потреб.

У сфері роботи з даними використання Agile-методологій дозволяє структурувати роботу над проєктами, що за своєю природою є невизначеними. Оскільки аналітичні проєкти відносно не часто мають чітко визначений результат на початку, ітеративний підхід забезпечує можливість швидкого експериментування,

перевірки гіпотез, гнучкої зміни цілей на основі нових даних та поступової побудови цінності. Водночас Agile під час роботи з даними та машинним навчанням вимагає певної адаптації, наприклад, традиційний спринт не завжди може завершитись "працюючим програмним забезпеченням", але може завершуватись аналітичним звітом або перевіреним прототипом моделі.

У практиці успішно застосовуються спеціалізовані адаптації Agile для аналітичних і науково-дослідних проєктів, зокрема, Data-Driven Scrum, Agile Data Science або комбінування Agile з іншими підходами, такими як CRISP-DM. Це дозволяє ефективно поєднувати структуровану ітеративність Agile із специфікою проєктів з дослідження даних, зберігаючи баланс між гнучкістю та дисципліною [7].

Таким чином, аналіз традиційних та гнучких підходів до управління проєктами демонструє необхідність адаптації методологій до контексту задач, рівня визначеності цілей, характеру даних та очікувань зацікавлених сторін. У випадку сфери роботи з даними та ML, саме гнучкі підходи, з урахуванням специфіки дослідницьких процесів, забезпечують найвищу ефективність реалізації проєктів.

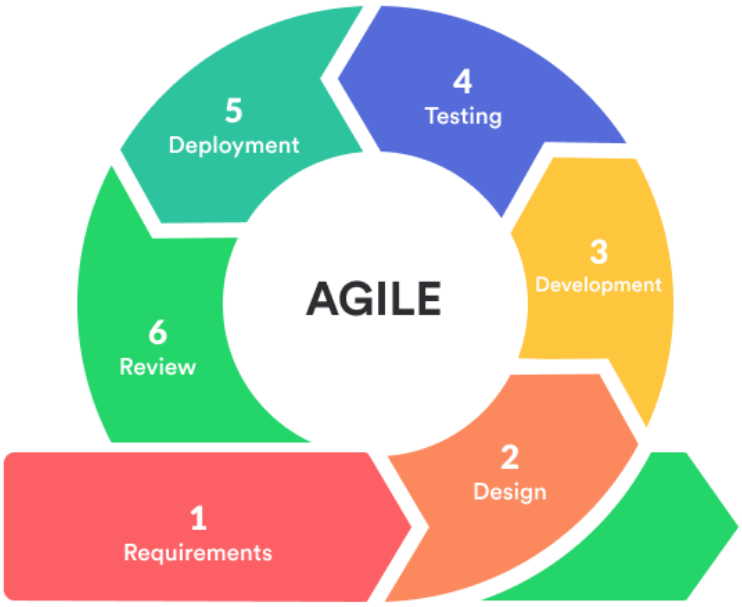

Рис. 1.1. Алгоритм розробки проєкту Agile підходу.

Проєкти в сфері роботи з даними характеризуються рядом унікальних особливостей, які кардинально відрізняють їх від традиційних IT-проєктів та

вимагають спеціалізованих підходів до управління. За визначенням Провоста та Фосетта, сфера машинного навчання та роботи з даними представляє собою набір принципів, що лежать в основі процесу видобування корисних знань з даних, де кожен етап вимагає особливого підходу до планування та контролю [8]. Експериментальний характер таких проєктів означає, що кінцевий результат не завжди можна точно передбачити на початку, оскільки він безпосередньо залежить від якості, повноти, структури наявних даних та ефективності обраних алгоритмічних підходів. Це створює унікальні виклики для проєктних менеджерів, які звикли працювати з чітко визначеними технічними вимогами та передбачуваними результатами розробки.

Ітеративність є ще однією ключовою характеристикою проєктів у сфері машинного навчання та роботи з даними, де процес розробки моделей вимагає постійного уточнення гіпотез та повторної обробки даних на основі отриманих проміжних результатів та нових інсайтів. Кожна ітерація може кардинально змінити розуміння проблеми та підходи до її вирішення, що вимагає від команди високого рівня гнучкості та готовності до швидких змін у напрямку досліджень. Міждисциплінарність успішної реалізації вимагає залучення експертів з різних галузей – статистики, програмування, предметної області, бізнес-аналізу, що створює додаткові складності в координації роботи, забезпеченні ефективної комунікації між учасниками команди та синхронізації їх зусиль.

Методологія Agile, яка довела свою високу ефективність у розробці програмного забезпечення, знайшла широке застосування в проєктах з машинного навчання та роботи з даними завдяки своїй природній відповідності експериментальному характеру аналітичних досліджень. Швабер та Сазерленд у своєму авторитетному керівництві підкреслюють критичну важливість адаптивності та ітеративного підходу, що особливо актуально для аналітичних проєктів, де кожен спринт може принести неочікувані відкриття та змінити траєкторію всього дослідження [13]. Застосування Agile принципів у сфері машинного навчання та роботи з даними передбачає організацію роботи через короткі ітерації тривалістю зазвичай два-три тижні, що дозволяє швидко тестувати гіпотези, отримувати цінний

зворотний зв'язок від зацікавлених сторін та оперативно коригувати напрямок досліджень відповідно до отриманих результатів.

Формування міжфункціональних команд стає критично важливим фактором успіху проєктів машинного навчання та роботи з даними, оскільки об'єднання науковців даних, дата-інженерів, галузевих експертів та представників бізнесу забезпечує комплексний підхід до вирішення завдань та мінімізує ризики неправильного трактування бізнес-вимог або технічних обмежень. Постійна комунікація через регулярні демонстрації проміжних результатів, ретроспективи та планування наступних спринтів дозволяє своєчасно коригувати напрямок досліджень та уникати витрат ресурсів на розробку неактуальних або неефективних рішень. Це особливо важливо в контексті розробки моделі утримання працівників, де глибоке розуміння специфіки HR-процесів, організаційної культури компанії та психологічних аспектів трудових відносин має вирішальне значення для створення практично корисного та ефективного аналітичного рішення.

Міжгалузевий стандартний процес для дослідження даних, відомий як CRISP-DM, залишається однією з найбільш популярних та широко визнаних методологій для структурування проєктів з аналізу даних у різних галузях промисловості та сферах застосування. Розроблена консорціумом провідних технологічних компаній у 1990-х роках за участі IBM, SPSS та NCR, ця методологія пройшла перевірку часом та довела свою практичну ефективність у сотнях реальних проєктів по всьому світу, від фінансового сектору до охорони здоров'я [7]. CRISP-DM визначає шість взаємопов'язаних та циклічних етапів, кожен з яких має свої специфічні цілі, конкретні завдання, необхідні ресурси та чіткі критерії успішного завершення.

Етап розуміння бізнесу є фундаментальним початком будь-якого проєкту у сфері машинного навчання та роботи з даними та передбачає глибокий аналіз бізнес-контексту, визначення всіх ключових стейкхолдерів та їх інтересів, формулювання чітких цілей з бізнес-перспективи та їх трансформацію в конкретні технічні завдання дослідження даних. Для проєкту розробки моделі утримання працівників цей етап включає детальний аналіз поточних HR-метрик організації, ідентифікацію та надавання пріоритетності факторів, що впливають на рівень утримання персоналу,

визначення цільових груп співробітників, та формулювання вимірюваних критеріїв успіху майбутньої прогностичної моделі. Етап розуміння даних включає систематичний збір початкових даних з усіх доступних джерел організації, їх детальне дослідження для виявлення внутрішньої структури, оцінку якості та повноти інформації, а також ідентифікацію потенційних проблем або обмежень, що можуть суттєво вплинути на якість та надійність майбутньої аналітичної моделі.

Підготовка даних традиційно є найбільш трудомістким та затратним по часу етапом проєкту з машинного навчання, який може займати від 60% до 80% загального часу проєкту та включає створення остаточного, очищеного та збагаченого набору даних для моделювання через комплексні процеси очищення від помилок та викидів, трансформації форматів та структур, інтеграції інформації з різнорідних джерел та створення нових інформативних змінних на основі наявних атрибутів. Етап моделювання передбачає науково обґрунтований вибір та практичне застосування найбільш підходящих методів машинного навчання для конкретного завдання, систематичне експериментування з різними алгоритмами та їх комбінаціями, ретельне налаштування гіперпараметрів для оптимізації продуктивності та створення ансамблів моделей для досягнення найкращих можливих результатів за обраними метриками якості.

Оцінка моделі включає всебічний аналіз її продуктивності на різних наборах даних, включаючи валідацію на незалежних тестових вибірках, перевірку відповідності отриманих результатів початковим бізнес-цілям проєкту, аналіз стабільності та надійності моделі в різних умовах експлуатації, а також підготовку детальних рекомендацій щодо практичного використання розробленого рішення. Завершальний етап розгортання охоплює організацію практичного використання результатів аналізу в реальному бізнес-середовищі, технічну інтеграцію моделі в існуючі інформаційні системи та бізнес-процеси організації, налаштування системи постійного моніторингу ефективності моделі та підготовку персоналу до роботи з новим аналітичним інструментом.

Процес виявлення знань у базах даних, відомий як KDD, запропонований Файядом та його колегами, представляє альтернативний більш детальний та технічно

орієнтований підхід до структурування проєктів аналізу даних, який виявляється особливо ефективним для складних багатоетапних аналітичних завдань з великими обсягами інформації [10]. На відміну від CRISP-DM, який збалансовано поєднує бізнес та технічні аспекти проєкту, KDD приділяє значно більшу увагу технічним деталям та математичним аспектам процесу видобування знань, включаючи додаткові етапи ретельної попередньої обробки даних, створення нових ознак (feature engineering) та поглибленої пост-обробки отриманих результатів для забезпечення їх практичної застосовності.

З іншого боку існує SEMMA методологія, розроблена та активно популяризована компанією SAS Institute, пропонує структурно відмінний погляд на організацію проєктів у сфері аналізу даних з особливим акцентом на застосуванні статистичних методів дослідження та розширеній візуалізації даних на всіх етапах роботи. Ця методологія структурно організована навколо п'яти послідовних основних етапів: вибірковий відбір репрезентативних зразків даних з генеральної сукупності, систематичне дослідження їх статистичних властивостей та внутрішньої структури, цілеспрямована модифікація та трансформація змінних для покращення їх аналітичної цінності, моделювання з використанням широкого спектра статистичних та машинних алгоритмів, та всебічна оцінка отриманих результатів з точки зору їх статистичної значущості та практичної застосовності. SEMMA виявляється особливо ефективною для наукових та дослідницьких проєктів, де статистична строгість, точність результатів та їх детальна інтерпретованість є критично важливими факторами успіху [10].

Варто зазначити, що сучасні проєкти у сфері роботи з даними все частіше інтегруються з принципами та найкращими практиками DevOps культури, що призводить до активного формування нової міждисциплінарної галузі під назвою MLOps або Machine Learning Operations. Цей інноваційний підхід забезпечує комплексну автоматизацію критично важливих процесів розробки, всебічного тестування, ретельної валідації та надійного розгортання моделей машинного навчання у продуктивному середовищі з високими вимогами до доступності та продуктивності [11]. MLOps як дисципліна включає систематичне версіонування

даних та моделей для забезпечення відтворюваності результатів, автоматизоване тестування якості вхідних даних та продуктивності розроблених моделей, впровадження принципів континуальної інтеграції та доставки спеціально адаптованих для ML-рішень, а також організацію постійного моніторингу продуктивності моделей в реальному часі з можливістю швидкого автоматичного виявлення та ефективного вирішення виникаючих проблем деградації якості або зміщення розподілу даних.

Застосування найкращих підходів до розробки та стратегічного планування проєктів у сфері аналізу даних дозволяє значно краще зрозуміти реальні потреби кінцевих користувачів системи, їх специфічні вимоги та обмеження, що в результаті дає можливість створювати більш релевантні, практично корисні та ефективні аналітичні рішення. Проте ключовим фактором успіху залишається правильно сформована та збалансована команда проєкту. Ефективна команда проєкту з розробки моделі машинного навчання зазвичай включає декілька ключових ролей, кожна з яких має свої чітко визначені специфічні обов'язки, необхідну експертизу та зони відповідальності в загальному процесі створення аналітичного рішення.

Розробка моделі утримання працівників IT-компанії є складним, багатоступеневим процесом, який вимагає чіткого планування, залучення міжфункціональної команди та використання сучасних методів аналітики. Для досягнення ефективного результату проєкт реалізується поетапно, із врахуванням як технічних, так і управлінських аспектів. Нижче наведено основні етапи реалізації такого проєкту [14]:

1. **Ініціація проєкту.** На цьому етапі визначаються загальна мета та задачі проєкту, обґрунтовується необхідність створення моделі утримання працівників як інструменту для зменшення плинності кадрів. Визначаються ключові зацікавлені сторони (HR-відділ, керівництво, аналітики), формулюється бачення проєкту та очікувані результати. Також оцінюються наявні ресурси та ризики.
2. **Планування проєкту.** Розробляється детальний план виконання робіт із розподілом відповідальності, термінами та контрольними точками.

Обираються інструменти та технології для збирання й аналізу даних (наприклад, Python, SQL, Power BI). Визначаються ключові показники ефективності (KPI) моделі, а також методи оцінювання результатів (точність прогнозу, коефіцієнт утримання тощо).

3. **Збір та підготовка даних.** Збирається історична інформація про співробітників компанії з внутрішніх систем (CRM, HRM, опитувальники). До релевантних даних можуть входити:

   – демографічні характеристики (вік, стать, стаж, рівень освіти);

   – професійні дані (посада, відділ, тип контракту, рівень зарплати);

   – показники продуктивності, участь у проєктах;

   – результати внутрішніх опитувань щодо задоволеності роботою;

   – дані про звільнення (дати, причини, тип звільнення).

Дані очищуються від пропусків, аномалій, проводиться кодування категоріальних змінних та нормалізація числових.

4. **Аналіз та інженерія ознак.** Виконується первинна аналітика для виявлення закономірностей, що можуть впливати на утримання працівників. Створюються нові ознаки (features), які покращують інформативність моделі, наприклад: рівень стресу за опитуваннями, стабільність кар'єрного зростання, взаємозв'язок з менеджером тощо.
5. **Розробка та навчання моделі.** Обираються відповідні алгоритми машинного навчання (наприклад, логістична регресія, дерева рішень, Random Forest або XGBoost) для побудови моделі, яка передбачає ймовірність звільнення кожного працівника. Дані розділяються на тренувальну та тестову вибірки. Застосовується крос-валідації, оптимізація гіперпараметрів, боротьба з дисбалансом класів.
6. **Оцінка ефективності моделі.** Модель тестується на відкладених даних. Оцінюються показники якості: Accuracy, Precision, Recall, F1-score, AUC-ROC. Аналізуються найвагоміші предиктори (важливість ознак) та моделюються різні сценарії утримання. За потреби проводиться донавчання або повторна побудова моделі.

7. **Візуалізація та презентація результатів.** Результати моделювання подаються у вигляді візуалізацій, звітів для прийняття управлінських рішень. Використовуються засоби візуалізації, як Power BI, Tableau або Python-бібліотек (Plotly, Seaborn). Готується презентація для HR-відділу та керівництва із висновками й рекомендаціями.
8. **Впровадження моделі.** Розробляються практичні кроки інтеграції моделі в HR-процеси: наприклад, включення її у систему раннього попередження, індивідуальні плани втримання працівників. Також можуть бути розроблені автоматизовані сповіщення для HR-партнерів про ризики звільнення.
9. **Моніторинг та підтримка.** Після запуску моделі важливо здійснювати постійний моніторинг її точності та актуальності. Створюються механізми збирання зворотного зв'язку, автоматизується оновлення даних та перенавчання моделі при зміні умов у компанії. Оцінюється реальний вплив моделі на скорочення плинності кадрів.

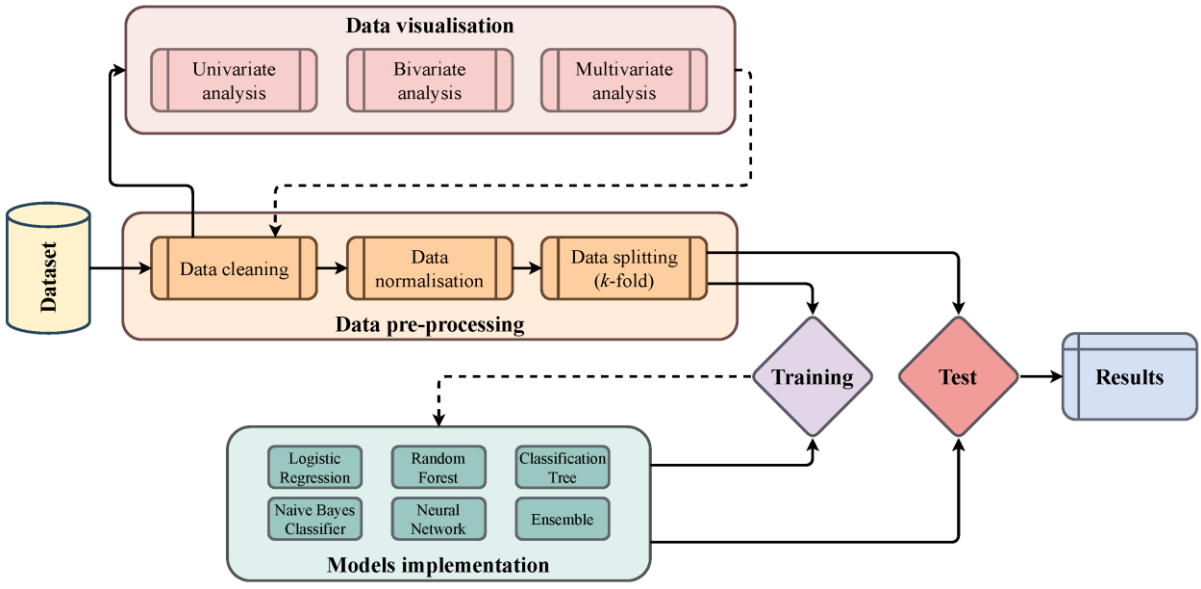

Рис. 1.2. Загальний алгоритм проєктів розробки моделі утримання працівників

**Джерело [15]*

Проєктний менеджер виконує функції координатора роботи всієї команди, забезпечує чітке розуміння бізнес-цілей та технічних вимог всіма учасниками проєкту, здійснює стратегічне планування пріоритетів розробки та розподіл ресурсів, організовує ефективну взаємодію із зацікавленими сторонами на всіх рівнях

організації, контролює дотримання встановлених термінів та бюджетних обмежень, а також забезпечує своєчасну доставку якісних результатів відповідно до узгоджених критеріїв успішності проєкту.

Оцінка ефективності та загальної успішності проєктів з розробки моделей машинного навчання вимагає застосування комплексного багаторівневого підходу, що збалансовано включає як технічні метрики якості та продуктивності розроблених моделей, так і стратегічні бізнес-показники практичної цінності та реального впливу отриманих результатів на ключові процеси організації. Технічні метрики охоплюють традиційні показники точності моделей такі як Accuracy, Precision, Recall, F1-score та AUC-ROC [12], показники швидкості навчання моделей та їх покращення в реальному часі, оцінки стабільності та надійності результатів при роботі з різними наборами даних та в мінливих умовах експлуатації, а також метрики споживання обчислювальних ресурсів та енергоефективності рішення.

Отже, бізнес-метрики зосереджуються на вимірюванні реального економічного та стратегічного впливу аналітичного рішення на діяльність організації через показники повернення інвестицій від впровадження проєкту з машинного навчання та роботи з даними, час до отримання вимірюваної бізнес-цінності від практичного використання моделі, конкретний позитивний вплив на ключові показники ефективності організації, рівень задоволеності кінцевих користувачів новими аналітичними можливостями, а також довгостроковий вплив на конкурентні переваги компанії на ринку. Тому заздалегідь необхідно ретельно обрати певний методологічний підхід відповідно до специфіки завдання та вимог, детально запланувати перебіг майбутнього проєкту з урахуванням всіх ризиків та обмежень, зібрати збалансовану команду з необхідними навичками, та визначити релевантні ключові показники ефективності для об'єктивної оцінки успішності проєкту на всіх етапах його реалізації.

1.2. Особливості та підходи до утримання працівників в IT-сфері

Інформаційні технології як галузь характеризуються унікальними особливостями, які кардинально впливають на підходи до управління персоналом та

утримання працівників. IT-сфера відрізняється від традиційних галузей економіки високим темпом розвитку, постійною потребою в інноваціях, специфічною організаційною культурою та особливими вимогами до кваліфікації персоналу.

Однією з ключових характеристик IT-індустрії є швидкість технологічних змін. Технології, які були актуальними ще кілька років тому, можуть втратити свою актуальність, що створює постійний тиск на працівників щодо оновлення своїх знань та навичок. Це формує особливу культуру безперервного навчання, де професійний розвиток стає не просто бажаним, а критично необхідним елементом кар'єри [15].

Специфіка проєктної роботи в IT також суттєво впливає на трудові відносини. На відміну від традиційних галузей, де робочі процеси можуть бути стандартизовані та передбачувані, IT-проєкти часто характеризуються високим рівнем невизначеності, творчим характером завдань та необхідністю швидкого реагування на зміни вимог клієнтів або ринку.

Глобальний характер IT-ринку створює унікальні можливості та виклики для управління персоналом. З одного боку, IT-спеціалісти мають доступ до міжнародних проєктів та можливості віддаленої роботи, що розширює їхні кар'єрні перспективи. З іншого боку, це створює додатковий тиск на роботодавців, оскільки конкуренція за таланти виходить далеко за межі локального ринку праці.

Демографічні особливості IT-працівників також відрізняються від інших галузей. Переважну частину IT-спеціалістів складають представники молодших поколінь, які мають специфічні очікування від роботи, цінують гнучкість, автономію та можливості для самореалізації. Ці працівники часто демонструють нижчу лояльність до конкретного роботодавця, але вищу прихильність до професійної спільноти та індустрії в цілому [16].

Дослідження показують, що фактори, які впливають на лояльність IT-працівників, мають свою специфіку порівняно з іншими галузями. Традиційні мотиватори, такі як стабільність роботи чи корпоративні пільги, можуть мати менший вплив на IT-спеціалістів, ніж на працівників інших сфер [16].

Професійний розвиток займає центральне місце серед факторів лояльності IT-працівників. Можливості для навчання нових технологій, участь у складних та

цікавих проєктах, доступ до сучасних інструментів та методологій розробки є критично важливими для утримання талантів. IT-спеціалісти високо цінують роботодавців, які інвестують в їхній професійний розвиток через тренінги, конференції, сертифікації та внутрішні програми навчання.

Опція віддаленої та гібридної роботи та гнучкість робочого процесу є ще одним важливим фактором лояльності. IT-працівники, особливо досвідчені спеціалісти, прагнуть мати контроль над своїми робочими процесами, вибором технологій та підходів до вирішення завдань. Мікроменеджмент та надмірний контроль можуть негативно впливати на їхню мотивацію та бажання залишатися в компанії. Баланс між роботою та особистим життям набуває особливого значення в контексті IT-індустрії, де поширені надурочні роботи та високий рівень стресу. Тому гнучкий графік роботи, наявність додаткових відпусток стають важливими факторами утримання працівників.

Можливості кар'єрного зростання в IT-сфері мають свої особливості. На відміну від традиційних ієрархічних структур, IT-спеціалісти часто цінують горизонтальне кар'єрне зростання, можливість переходу між різними проєктами, командами чи технологічними напрямками. Експертні кар'єрні шляхи, які дозволяють розвиватися як технічний спеціаліст без обов'язкового переходу до управлінських позицій, особливо цінуються в IT-середовищі [16].

Організаційна культура відіграє особливу роль в IT-компаніях. Культура інновацій, відкритості до експериментів, толерантності до помилок та заохочення ініціативи є критично важливою для залучення та утримання IT-талантів. Працівники цінують середовище, де їхні ідеї можуть бути почуті та реалізовані, де заохочується творчий підхід до вирішення проблем.

Компенсаційні пакети в IT-сфері також мають свою специфіку. Окрім базової заробітної плати, IT-працівники часто очікують отримання акцій компанії, бонусів за результати проєктів, компенсації витрат на професійний розвиток. Прозорість та справедливість системи винагороди є важливими факторами довіри до роботодавця [16].

Соціальні зв'язки та командна робота, незважаючи на технічну природу роботи, відіграють значну роль у формуванні лояльності ІТ-працівників. Можливість працювати з талановитими колегами, обмінюватися досвідом, брати участь у професійних спільнотах, технічних вебінарах сприяє формуванню зв'язку з компанією та іншими колегами.

Плинність кадрів в ІТ-сфері традиційно є вищою порівняно з іншими галузями економіки. Дослідження показують, що середній показник плинності в ІТ-компаніях може досягати 20-30% на рік, що перевищує середньогалузеві показники 5-15% на рік [17].

Недостатні можливості професійного розвитку є однією з основних причин звільнення ІТ-спеціалістів. В умовах швидкого розвитку технологій, працівники, які не мають можливості оновлювати свої навички, швидко втрачають конкурентоспроможність на ринку праці. Компанії, які не інвестують в розвиток своїх співробітників, ризикують їх втратити. Робота зі застарілими технологіями може сприйматися як гальмування професійного розвитку та призводити до пошуку більш технологічно прогресивного роботодавця.

Таким чином відсутність чітких перспектив кар'єрного зростання є ще однією поширеною причиною звільнень. ІТ-спеціалісти, особливо амбітні та талановиті, прагнуть бачити чіткі шляхи свого професійного розвитку. Компанії з плоскою організаційною структурою або обмеженими можливостями просування можуть зіткнутися з проблемою утримання перспективних кадрів.

Концепція бренду роботодавця набуває особливого значення в умовах високої конкуренції за ІТ-таланти. Розробка сильного бренду роботодавця включає формування та просування унікальної ціннісної пропозиції для працівників, активну присутність у професійних спільнотах, демонстрацію корпоративної культури та цінностей.

З іншого боку низька компенсація залишається значним фактором плинності, особливо в умовах високого попиту на ІТ-спеціалістів. Динамічність ринку праці в ІТ-сфері призводить до зростання рівня заробітних плат, і компанії, які не відслідковують ринкові тенденції, можуть втратити своїх найкращих працівників

[18]. Концепція загальної мотивації передбачає комплексний підхід до винагороди працівників, що включає не лише грошову компенсацію, але й можливості розвитку, визнання досягнень, баланс роботи та особистого життя, сприятливе робоче середовище. Цей підхід особливо ефективний в IT-сфері, де працівники цінують різноманітні форми винагороди.

Незадовільне управління та лідерство негативно впливають на рівень плинності в IT-командах. Неефективне управління проєктами, відсутність чіткого бачення цілей, конфлікти в командах можуть призводити до фрустрації та бажання змінити роботу. Особливо це стосується ситуацій, коли технічні рішення приймаються людьми без достатньої технічної компетентності. Токсична організаційна культура може стати критичним фактором плинності. Це включає відсутність довіри між співробітниками та керівництвом, несправедливе ставлення, відсутність визнання досягнень, надмірний контроль. IT-спеціалісти, маючи широкі можливості вибору роботи, не схильні терпіти негативне робоче середовище [18].

Як було наведено вище, перевантаження та виснаження є поширеною проблемою в IT-індустрії. Постійні дедлайни, необхідність працювати понаднормово, високий рівень стресу можуть призводити до професійного вигорання та бажання змінити сферу діяльності або роботодавця. Невідповідність очікувань реальності може стати причиною швидкого звільнення нових співробітників. Це особливо актуально для випускників університетів або спеціалістів, які переходять з інших галузей. Неадекватне представлення вакансії, реальних обов'язків чи корпоративної культури під час процесу найму може призвести до розчарування та швидкого звільнення.

Розвиток теорії та практики управління персоналом в IT-сфері призвів до формування специфічних підходів до утримання працівників. Ці підходи враховують особливості IT-індустрії та специфічні потреби IT-спеціалістів.

Орієнтований на досвід працівника підхід набуває все більшої популярності в IT-компаніях. Цей підхід передбачає комплексне управління всіма аспектами взаємодії працівника з компанією - від процесу найму до завершення трудових

відносин. Особлива увага приділяється створенню позитивного досвіду на всіх етапах робочого життя співробітника.

Персоналізований підхід до управління утриманням враховує індивідуальні потреби та мотивації кожного працівника. Це включає розробку індивідуальних планів розвитку, гнучких умов праці, персоналізованих пакетів винагороди. Використання даних та аналітики дозволяє HR-службам краще розуміти потреби конкретних співробітників та адаптувати свої підходи відповідно

Agile-підходи до управління персоналом запозичують принципи гнучкої методології розробки програмного забезпечення. Це включає регулярний зворотний зв'язок, швидке реагування на зміни, ітеративний підхід до розвитку процесів HR. Такі методи дозволяють швидко адаптуватися до змін потреб працівників та ринкових умов [19].

Одним із ключових підходів до зниження плинності кадрів є впровадження системного HR-аналітичного підходу. Суть цього підходу полягає в глибокому аналізі даних про працівників, включаючи історію зайнятості, рівень задоволеності, динаміку заробітної плати, участь у внутрішніх ініціативах тощо. Використання таких інструментів, як моделювання на основі прогнозних алгоритмів, дозволяє виявляти ризикові групи працівників, схильних до звільнення, та вчасно реагувати шляхом індивідуалізації управлінських рішень. Аналітика утримання дозволяє не лише оцінити поточний стан задоволеності та лояльності персоналу, а й формувати довгострокову стратегію зменшення кадрових втрат.

Поряд з аналітичними підходами, важливою складовою є стратегічне управління талантами. Цей підхід передбачає планомірне виявлення, розвиток і збереження ключових працівників, які мають потенціал до лідерства або є критично важливими для проєктів компанії. Ефективне управління талантами базується на концепції життєвого циклу робітника, що охоплює всі етапи взаємодії працівника з компанією – від найму до виходу з організації. На кожному етапі життєвого циклу працівника застосовуються цільові заходи щодо підвищення мотивації, розвитку навичок та забезпечення відчуття значущості у межах команди.

Ще одним важливим підходом до утримання працівників є побудова позитивної організаційної культури. Успішні ІТ-компанії не лише забезпечують своїм працівникам конкурентну оплату праці, а й формують ціннісне середовище, в якому поважають індивідуальність, стимулюють ініціативу та забезпечують зворотний зв'язок. Організаційна культура, орієнтована на розвиток, відкритість до змін та підтримку балансу між роботою та особистим життям, створює умови для довгострокової співпраці. Працівники, які відчувають емоційний зв'язок із компанією, демонструють вищий рівень продуктивності та нижчу схильність до змін місця роботи.

Сучасні підходи до мотивації працівників також змінюються в бік індивідуалізації. Традиційні підходи, що базуються виключно на матеріальних винагородах, поступаються місцем комплексним мотиваційним стратегіям, які враховують особистісні цілі, кар'єрні очікування, рівень автономії та потребу в реалізації. Підвищення гнучкості в управлінні працівниками, включаючи можливість гібридного графіка, внутрішньої ротації або участі в міжфункціональних проєктах, дозволяє утримувати таланти шляхом задоволення їх професійних амбіцій.

У цьому контексті особливе значення відіграє роль HR-підрозділу, який трансформується з адміністративної функції на стратегічного партнера. HR-фахівці сьогодні виступають як прискорювачі змін, впроваджуючи інструменти аналітики, управління продуктивністю, зворотного зв'язку та розвитку лідерських якостей. Вони активно співпрацюють з менеджерами проєктів для забезпечення єдиної політики утримання персоналу, адаптованої до потреб конкретної команди та бізнес-напряму.

Проєктний менеджмент в ІТ-компаніях також відіграє ключову роль у формуванні умов, що сприяють утриманню працівників. По-перше, саме проєктні менеджери щодня взаємодіють з командами та можуть оперативно виявляти сигнали незадоволеності або ризику вигорання. По-друге, в межах гнучких методологій управління (Agile, Scrum, Kanban) активно застосовуються практики регулярного зворотного зв'язку, командних ретроспектив, планування спринтів, що дозволяє забезпечити прозору комунікацію, своєчасне реагування на проблеми та залученість кожного члена команди в процес ухвалення рішень.

Особливо важливою є адаптація стилю управління проєктами під специфіку команди та індивідуальні особливості працівників. Практика лідерства дозволяє менеджерам формувати довірливі стосунки з командою, підтримуючи особистісний та професійний розвиток кожного учасника. Такий підхід сприяє формуванню середовища психологічної безпеки, де працівники не бояться висловлювати ідеї, ділитися проблемами та брати відповідальність.

Крім того, важливо забезпечувати кар'єрне моделювання в межах проєктів – тобто створення можливостей для горизонтального або вертикального зростання без необхідності змінювати місце роботи. Це може включати створення внутрішніх академій, менторських програм, доступу до сертифікацій або участі в міжпроєктних ініціативах. Такий підхід дозволяє поєднати розвиток проєктного потенціалу з індивідуальними потребами працівника, зменшуючи ризик звільнень через відчуття застою чи невизнання [19].

Таким чином, утримання працівників в IT-сфері є багаторівневим завданням, що вимагає комплексного підходу, в якому аналітика, стратегічне управління талантами, організаційна культура та лідерські практики поєднуються в єдину систему. Роль HR та менеджерів проєктів у цьому процесі є ключовою, оскільки саме від їхньої здатності до співпраці, адаптації та впровадження гнучких інструментів управління залежить не лише стабільність команди, а й конкурентоспроможність компанії загалом.

Завдяки усім наведеним вище факторам використання технологій для управління персоналом стає все більш важливим в IT-компаніях. HR-аналітика дозволяє виявляти ранні сигнали можливого звільнення працівників, аналізувати фактори, що впливають на задоволеність роботою, та приймати управлінські рішення.

Платформи для управління талантами інтегрують різні HR-процеси та дозволяють створити єдиний простір для взаємодії з працівниками. Це включає системи управління продуктивністю, платформи для навчання та розвитку, інструменти для зворотного зв'язку та визнання досягнень.

Ще більше змінюють процес взаємодії з працівниками сучасні тренди – штучний інтелект та машинне навчання. Вони знаходять все більше застосування в

HR-процесах. Це включає автоматизацію рутинних завдань, персоналізацію досвіду працівників, прогнозування плинності кадрів, оптимізацію процесів найму та розвитку [20]. Автоматизація та штучний інтелект створюють нові можливості, але також породжують побоювання щодо майбутнього робочих місць. Компанії повинні активно управляти цими змінами, забезпечуючи перекваліфікацію працівників та чітко доносячи свої наміри щодо використання нових технологій.

Мобільні та браузерні HR-застосунки дозволяють працівникам отримувати доступ до сервісів компанії в будь-який час та в будь-якому місці. Це особливо актуально для IT-спеціалістів, які часто працюють віддалено або мають гнучкий графік роботи.

Також аналіз міжнародного досвіду показує різноманітність підходів до управління утриманням IT-працівників в різних країнах та регіонах. Скандинавські країни демонструють успішні практики створення інклюзивного та підтримуючого робочого середовища, де високо цінуються баланс роботи та особистого життя, рівність можливостей та соціальна відповідальність. Досвід Кремнієвої долини показує ефективність агресивних стратегій залучення та утримання талантів, включаючи щедрі компенсаційні пакети, унікальні пільги та можливості, інноваційну корпоративну культуру. Однак такі підходи можуть бути не завжди застосовні в інших економічних умовах [21].

Європейські IT-компанії часто акцентують увагу на стабільності, професійному розвитку та дотриманні балансу роботи та особистого життя. Це включає законодавчо закріплені права працівників, розвинені системи соціального захисту, культуру поваги до особистого часу.

З іншого боку Азіатські IT-компанії демонструють унікальні підходи, які поєднують традиційні цінності з сучасними HR-практиками. Це включає акцент на довгострокових відносинах з працівниками, інвестиції в масштабні програми розвитку, створення сильної корпоративної ідентичності [21].

Сучасні виклики в управлінні утриманням IT-працівників пов'язані з кількома ключовими тенденціями. Віддалена робота, яка стала поширеною після пандемії

COVID-19, створює нові виклики для підтримання командного духу, корпоративної культури та ефективної комунікації.

Зростаюча важливість цілеспрямованої кар'єри означає, що IT-спеціалісти все частіше шукають роботу, яка має значення та сприяє вирішенню важливих соціальних або екологічних проблем. Компанії повинні чітко артикулювати свою місію та показувати, як робота кожного співробітника сприяє досягненню більших цілей.

Різноманітність вікових груп в робочому колективі створює додаткові виклики, оскільки різні покоління мають різні очікування від роботи, комунікаційних стилів та мотиваційних факторів. HR-стратегії повинні враховувати ці відмінності та забезпечувати інклюзивне середовище для всіх вікових груп.

Зростання значення ментального здоров’я та самопочуття на роботі стає все більш важливим фактором утримання працівників. IT-компанії все частіше впроваджують програми підтримки психічного здоров'я, стресменеджменту та профілактики професійного вигорання, наприклад, через спеціалізовані курси та вебінари.

Успішне управління утриманням працівників передбачає також інтеграцію HR та проєктних стратегій через єдину систему KPI, в яку включаються не лише показники ефективності, а й метрики залученості, задоволеності та індексу утримання. Це дозволяє не лише кількісно вимірювати ефективність управлінських практик, а й формувати культуру відповідальності за людський капітал як на рівні HR, так і на рівні кожного менеджера [22].

Індекс лояльності співробітників (Employee Net Promoter Score) адаптує концепцію індексування для оцінки лояльності працівників. Цей показник вимірює готовність співробітників рекомендувати свою компанію як місце роботи. Для IT-компаній це особливо важливо, оскільки рекомендації колег є одним з найефективніших каналів залучення нових талантів. Зазвичай він обраховується за допомогою проходження працівниками спеціалізованих анонімних опитувальників.

Термін від працевлаштування до продуктивної роботи вимірює час, необхідний новому співробітнику для досягнення продуктивності інших колег на аналогічній

посаді. В IT-сфері цей показник особливо важливий через складність технічних систем, тривалого процесу онбордингу на проєкти та необхідність глибокого розуміння бізнес-процесів. Швидша адаптація часто корелює з вищим рівнем задоволеності роботою [22].

Коефіцієнт залученості працівника також вимірюється через регулярні опитування та вимірюють рівень залученості працівників до корпоративної культури всередині компанії та лояльності до роботодавця. Високий рівень залученості часто передує високим результатам утримання.

Вартість найму нового співробітника дозволяє оцінити ефективність процесів найму та їх вплив на подальше утримання працівників. В умовах високої конкуренції за IT-таланти ці метрики допомагають оптимізувати інвестиції в залучення персоналу [22].

Можна дійти висновку, що управління утриманням працівників в IT-сфері характеризується значною специфікою, яка вимагає адаптації традиційних HR-підходів до особливостей галузі. Ключовими факторами успіху є розуміння унікальних мотивацій IT-спеціалістів, використання сучасних технологій та методологій управління персоналом, а також постійна адаптація до змін ринкових умов. Ефективні стратегії утримання IT-працівників повинні бути комплексними та включати елементи професійного розвитку, конкурентної компенсації, сприятливої організаційної культури та можливостей кар'єрного зростання. Особлива увага повинна приділятися персоналізації підходів та використанню даних для прийняття рішень.

Майбутній розвиток підходів до управління утриманням IT-працівників буде пов'язаний з подальшою цифровізацією та автоматизацією HR-процесів, зростаючим значенням досвідченості співробітників та необхідністю адаптації до нових форм організації праці.

1.3. Огляд методів та підходів до машинного навчання для прогнозування відтоку працівників

Машинне навчання представляє собою галузь штучного інтелекту, що дозволяє алгоритмам навчатися на наперед зібраному і очищеному наборі даних, без явного програмування для виконання конкретних завдань. У сучасному корпоративному середовищі ця технологія знаходить широке застосування в різних аспектах управління персоналом, особливо в задачах прогнозування поведінки працівників та оптимізації кадрових процесів.

Застосування методів машинного навчання для прогнозування відтоку працівників базується на здатності алгоритмів виявляти складні залежності та приховані закономірності в даних про співробітників. Це дозволяє HR-департаментам та керівництву компаній створювати персоналізовані стратегії утримання працівників, які враховують індивідуальні особливості кожного співробітника та фактори ризику, що можуть призвести до звільнення [24].

Ефективність машинного навчання в задачах прогнозування відтоку працівників обумовлена декількома ключовими факторами. По-перше, сучасні IT-компанії генерують величезні обсяги структурованих та неструктурованих даних про своїх співробітників, включаючи демографічну інформацію, дані про продуктивність, результати опитувань задоволеності, історію кар'єрного зростання, відгуки про роботу в команді тощо. По-друге, традиційні статистичні методи часто виявляються недостатніми для аналізу таких складних багатовимірних даних, тоді як алгоритми машинного навчання здатні ефективно обробляти великі набори даних та виявляти нелінійні взаємозв'язки між змінними [24].

Особливої актуальності машинне навчання набуває в IT-індустрії, де рівень плинності кадрів традиційно є високим через специфіку галузі, інтенсивну конкуренцію за талановитих фахівців та динамічний характер технологічного розвитку. За даними досліджень, середній рівень відтоку в IT-компаніях складає від 15% до 25% річно, що значно перевищує показники в інших галузях економіки. Такі

високі показники плинності створюють значні витрати на рекрутинг, навчання нових співробітників та втрату інтелектуального капіталу [16].

Існують три основних типи машинного навчання, кожен з яких має свої особливості та області застосування в контексті прогнозування відтоку працівників: навчання з учителя, навчання без учителя та підсилене навчання.

Навчання з учителем є найбільш поширеним підходом для розв'язання задач прогнозування відтоку працівників. Цей тип навчання характеризується використанням історичних даних, де для кожного спостереження відомий результат – залишився працівник у компанії чи звільнився. Алгоритми навчання з учителем аналізують взаємозв'язки між характеристиками працівників (незалежні змінні) та фактом їх звільнення (залежна змінна), після чого будують прогностичні моделі, здатні передбачати ймовірність відтоку нових співробітників [25].

Перевагою навчання з учителем є можливість отримання більш точних та надійних результатів, оскільки алгоритм навчається на основі фактичних даних про попередні випадки звільнень. Однак цей підхід має й певні обмеження, зокрема потребу в значних обсягах якісних історичних даних та можливість виникнення проблем з узагальненням моделі на нові дані, якщо умови в компанії істотно змінилися порівняно з періодом навчання.

Навчання без учителя використовується в ситуаціях, коли відсутня чітка інформація про результат або коли необхідно виявити приховані структури в даних про працівників. Основними застосуваннями цього типу навчання в HR-аналітиці є кластеризація співробітників за схожими характеристиками, виявлення закономірностей поведінки та зменшення розмірності даних для підготовки до подальшого аналізу [25].

Кластерний аналіз дозволяє виділити групи працівників зі схожими характеристиками, поведінкою та ризиками звільнення. Це забезпечує можливість розробки диференційованих стратегій утримання для різних категорій співробітників та більш ефективного розподілу ресурсів HR-департаменту.

Підсилене навчання, хоча й менш поширене в HR-аналітиці, може застосовуватися для оптимізації довгострокових стратегій утримання працівників.

Цей підхід базується на взаємодії агента з середовищем через систему винагород та покарань, що дозволяє поступово удосконалювати стратегії прийняття рішень. У контексті управління персоналом підсилене навчання може використовуватися для оптимізації процесів підвищення заробітної плати, планування кар'єрного зростання та розподілу бонусів з метою максимізації утримання ключових співробітників [26].

Ефективність будь-якої моделі машинного навчання критично залежить від якості вхідних даних. У контексті прогнозування відтоку працівників процес підготовки даних включає декілька ключових етапів: збір даних, їх очищення, перетворення та відбір факторних ознак.

Збір даних про працівників може здійснюватися з різноманітних джерел. Традиційними джерелами є кадрові інформаційні системи, які містять базову демографічну інформацію, дані про освіту, стаж роботи, посадові обов'язки та історію змін у кар'єрі. Додатково важливими є дані з систем управління продуктивністю, які фіксують результати оцінки ефективності, досягнення цілей, участь у проєктах та професійному розвитку.

Сучасні IT-компанії також активно використовують дані з корпоративних платформ комунікації, систем контролю версій коду, платформ для управління проєктами та інших цифрових інструментів. Ці джерела надають цінну інформацію про рівень залученості працівників, інтенсивність співпраці з колегами, участь у корпоративних заходах та загальну активність у робочому середовищі.

Регулярні опитування працівників представляють ще одне важливе джерело даних. Опитування задоволеності роботою, оцінки корпоративної культури, відгуки про керівництво та колег дозволяють кількісно оцінити суб'єктивні аспекти трудового досвіду, які часто є ключовими факторами схильності до звільнення.

Оцінка якості даних здійснюється за допомогою різноманітних статистичних методів. Дескриптивна статистика дозволяє отримати загальне уявлення про розподіл значень змінних, виявити потенційні проблеми та аномалії. Кореляційний аналіз допомагає виявити взаємозв'язки між різними характеристиками працівників та ідентифікувати потенційно важливі фактори відтоку. Регресійний аналіз може

використовуватися для попередньої оцінки впливу окремих факторів на ймовірність звільнення [27].

Існують різні підходи до відбору ознак. Статистичні методи, такі як аналіз головних компонент (PCA), дозволяють зменшити розмірність даних шляхом створення нових змінних, які є лінійними комбінаціями оригінальних ознак. Методи, засновані на важливості ознак, використовують алгоритми машинного навчання для оцінки внеску кожної змінної в прогнозування цільової змінної [27].

Зазвичай задача прогнозування відтоку працівників формулюється як бінарна класифікація, де необхідно визначити, до якого з двох класів належить кожен співробітник: "залишається в компанії" або "звільняється". Для розв'язання цієї задачі застосовуються різноманітні класифікаційні алгоритми машинного навчання, кожен з яких має свої особливості, переваги та обмеження.

1.1.1. Логістична регресія (Logistic Regression)

Логістична регресія є одним з найпоширеніших та найбільш інтерпретованих методів для бінарної класифікації. У контексті прогнозування відтоку працівників цей алгоритм моделює ймовірність звільнення як функцію від характеристик співробітника.

Математичну основу логістичної регресії можна описати наступними рівняннями [28]:

$$p(y = 1 \mid x_1,\ldots,x_n) = f(y) \tag{1.1}$$

$$f(y) = \frac{1}{(1+e^{-y})} \tag{1.2}$$

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n \tag{1.3}$$

, де:

- y - цільова змінна для кожного індивідуума j (працівника в моделюванні відтоку), y - бінарна мітка класу (0 або 1);
- β_0 є константою;
- β_1 вага, що надається конкретній змінній, пов'язаній з кожним працівником j (j=1,.... m);

$x_1,\ldots,x_n$ це факторні (незалежні) змінні для кожного працівника j, на основі яких потрібно спрогнозувати y.

Основною перевагою логістичної регресії є її інтерпретованість. Коефіцієнти моделі можуть бути безпосередньо інтерпретовані як логарифми відношень ймовірностей, що дозволяє HR-фахівцям зрозуміти, які фактори та в якій мірі впливають на ймовірність звільнення працівників. Наприклад, позитивний коефіцієнт при змінній "кількість відпрацьованих років" вказує на те, що зі збільшенням стажу зростає ймовірність звільнення, що може свідчити про професійне вигорання або пошук нових можливостей досвідченими співробітниками.

Логістична регресія також вимагає відносно невеликих обчислювальних ресурсів та швидко навчається навіть на великих наборах даних. Модель стійка до викидів та не потребує складного налаштування гіперпараметрів, що робить її привабливим вибором для початкового дослідження даних.

Однак логістична регресія має певні обмеження. Вона припускає лінійний зв'язок між логарифмом відношення шансів та незалежними змінними, що може бути неадекватним для моделювання складних нелінійних залежностей в HR-даних. Крім того, алгоритм чутливий до мультиколінеарності між незалежними змінними, що може призвести до нестабільності коефіцієнтів та помилкових висновків про важливість окремих факторів.

1.1.2. Дерева рішень (Decision Trees)

Дерева рішень представляють собою один з найбільш інтуїтивно зрозумілих алгоритмів машинного навчання, який створює деревоподібну структуру правил для класифікації об'єктів. У контексті прогнозування відтоку працівників дерево рішень будує ієрархію умов, які поступово розділяють співробітників на групи з різною ймовірністю звільнення [29].

Процес побудови дерева рішень починається з кореневого вузла, який представляє всю множину працівників. Алгоритм вибирає характеристику та її порогове значення, які найкращим чином розділяють дані на дві підгрупи з різними рівнями ризику звільнення. Цей процес повторюється для кожної підгрупи до досягнення певного критерію зупинки.

Критерієм для вибору найкращого розбиття зазвичай служить інформаційний приріст або коефіцієнт Джині. Інформаційний приріст вимірює зменшення ентропії (міри невизначеності) після розбиття, тоді як коефіцієнт Джині оцінює ймовірність неправильної класифікації випадково вибраного елемента.

Основною перевагою дерев рішень є їх високу інтерпретованість. Результат роботи алгоритму може бути легко візуалізований у вигляді схеми, яку можуть зрозуміти навіть фахівці без глибоких знань в області машинного навчання. Це особливо важливо для HR-департаментів, де рішення повинні бути обґрунтованими та зрозумілими для керівництва компанії [29].

Дерева рішень також мають ряд технічних переваг. Вони не потребують попередніх припущень щодо розподілу даних, можуть ефективно обробляти як числові, так і категоріальні змінні, та природно враховують взаємодії між різними характеристиками працівників. Алгоритм стійкий до викидів та може працювати з пропущеними значеннями без докової попередньої обробки. Однак дерева рішень мають схильність до перенавчання, особливо при роботі з великими та складними наборами даних. Глибокі дерева можуть створювати занадто специфічні правила, які добре працюють на навчальних даних, але погано узагальнюються на нові спостереження. Для вирішення цієї проблеми застосовуються техніки обрізання дерева, які видаляють гілки з найвищою ймовірністю помилки [29].

Іншим недоліком є нестабільність дерев рішень, а саме невеликі зміни в навчальних даних можуть призвести до істотних змін у структурі дерева. Це ускладнює інтерпретацію результатів та може знижувати довіру до моделі з боку кінцевого замовника.

1.1.3. Метод опорних векторів (Support Vector Machine)

Метод опорних векторів (SVM) є потужним алгоритмом, який може ефективно розв'язувати як лінійні, так і нелінійні задачі класифікації. Основна ідея методу полягає у знаходженні оптимальної гіперплощини, яка найкращим чином розділяє працівників різних класів у багатовимірному просторі характеристик [30].

У випадку лінійно розділюваних даних SVM знаходить гіперплощину, яка максимізує відстань до найближчих точок кожного класу. Ці найближчі точки

називаються опорними векторами і визначають положення поділяючої межі. Максимізація відстані до опорних векторів забезпечує кращу здатність моделі до узагальнення на нові дані.

У контексті прогнозування відтоку працівників SVM може ефективно виявляти складні нелінійні залежності між характеристиками співробітників та їх схильністю до звільнення. Наприклад, алгоритм може виявити, що комбінація певного рівня зарплати, стажу роботи та результатів оцінки ефективності створює специфічний профіль ризику, який не може бути адекватно описаний простими лінійними моделями.

Перевагами SVM є висока точність класифікації, особливо при роботі з даними високої розмірності, та ефективність у випадках, коли кількість ознак перевищує кількість спостережень. Алгоритм також стійкий до перенавчання завдяки принципу максимізації маржі.

Основними недоліками SVM є складність інтерпретації результатів, особливо при використанні нелінійних ядрових функцій, та чутливість до масштабування даних. Іншим ускладненням є потреба налаштування гіперпараметрів моделі.

1.1.4. Наївний алгоритм Баєса (Naive Bayes)

Наївний алгоритм Баєса базується на теоремі Баєса та припущенні про умовну незалежність ознак. Цей алгоритм часто показує хороші результати в задачах класифікації та є особливо корисним при роботі з великими наборами даних [31].

Ймовірність допомагає передбачити настання події з усіх можливих результатів. Математичне рівняння ймовірності виглядає наступним чином: Ймовірність події = Кількість сприятливих подій / Загальна кількість результатів 0 < = ймовірність події < = 1. Сприятливий результат означає подію, яка є результатом ймовірності. Ймовірність завжди знаходиться в діапазоні від 0 до 1, де 0 означає відсутність ймовірності того, що подія відбудеться, а 1 означає високу ймовірність того, що подія відбудеться [31].

Теорія Байєса працює над тим, щоб прийти до гіпотези (H) на основі певного набору доказів (E). Вона стосується двох речей: ймовірності гіпотези до доказів P(H) та ймовірності після доказів P(H|E). Теорія Байєса пояснюється наступним рівнянням:

$$P(H|E) = (P(E|H\} * P(H))/P(E) \quad (1.4)$$

, де

- P(H|E) показує, як часто подія H відбувається, коли відбувається подія E першою;
- P(E|H) показує, як часто подія E відбувається, коли подія H відбувається першою;
- P(H) показує ймовірність того, що подія X відбудеться сама по собі;
- P(E) - ймовірність того, що подія Y відбудеться сама по собі.

Правило Байєса – це метод визначення P(H|E) з P(E|H). Тобто воно надає вам спосіб обчислення ймовірності гіпотези на основі наданих спостережень.

Алгоритм обчислює ймовірність належності працівника до кожного класу на основі його характеристик та присвоює йому клас з найвищою ймовірністю. Припущення про незалежність дозволяє розкласти ймовірність P(характеристики | клас) як добуток ймовірностей окремих ознак, що значно спрощує обчислення.

У контексті HR-аналітики цей алгоритм може бути особливо корисним при роботі з категоріальними даними, такими як рівень освіти, тип посади, департамент, або результати внутрішніх опитувань. Алгоритм також ефективно працює з текстовими даними, що робить його придатним для аналізу відгуків співробітників або відкритих відповідей в опитуваннях [32].

Основними перевагами наївного алгоритму Баєса є швидкість навчання та прогнозування, простота реалізації та хороша продуктивність при обмеженій кількості навчальних даних. Алгоритм також природно надає ймовірнісні оцінки належності до класу, що може бути корисним для ранжирування працівників за рівнем ризику звільнення.

Головним недоліком є нереалістичність припущення про незалежність ознак, яке майже ніколи не виконується в даних реального бізнесу. Наприклад, рівень заробітної плати часто корелює з посадою, стажем роботи та рівнем освіти. Порушення цього припущення може призвести до неточних оцінок ймовірностей та погіршення якості класифікації.

1.1.5. Алгоритм k-найближчих сусідів (k-NN)

Алгоритм k-найближчих сусідів (k-NN) є простим, але ефективним методом класифікації, який базується на принципі подібності об'єктів. Для класифікації нового працівника алгоритм знаходить k найбільш подібних співробітників у навчальному наборі (найближчих сусідів) та присвоює новому об'єкту клас, який найчастіше зустрічається серед цих сусідів [33].

Ключовим аспектом алгоритму є вибір метрики відстані для вимірювання подібності між працівниками. Найпоширенішими є евклідова відстань для числових ознак та відстань Хеммінга для категоріальних змінних. При роботі зі змішаними типами даних може використовуватися метрика Гауера, яка поєднує різні типи відстаней.

Параметр k (кількість сусідів) є критично важливим для продуктивності алгоритму. Малі значення k можуть призводити до перенавчання та високої чутливості до шуму в даних, тоді як великі значення можуть згладжувати локальні закономірності та знижувати точність класифікації. Зазвичай оптимальне значення k визначається за допомогою крос-валідації.

У контексті прогнозування відтоку працівників k-NN може бути корисним для ідентифікації співробітників з подібними профілями ризику. Наприклад, якщо декілька працівників з певними характеристиками (вік, посада, стаж, рівень задоволеності) звільнилися протягом останнього року, алгоритм буде класифікувати нових співробітників з подібними характеристиками як групу високого ризику [34].

Перевагами k-NN є простота концепції та реалізації, відсутність припущень щодо розподілу даних та здатність моделювати складні нелінійні межі класів. Алгоритм також не потребує періоду навчання – всі обчислення виконуються на етапі прогнозування.

Основними недоліками є висока обчислювальна складність прогнозування, адже необхідно обчислювати відстані до всіх точок навчального набору, погіршення продуктивності при великій кількості ознак. Алгоритм також чутливий до масштабування ознак та може давати нестабільні результати при наявності нерівномірно розподілених класів.

1.1.6. Випадковий ліс (Random Forest)

Випадковий ліс (Random Forest) представляє собою ансамблевий метод, який комбінує прогнози множини дерев рішень для отримання більш стабільних та точних результатів. Алгоритм будує велику кількість дерев на різних підвибірках навчальних даних та різних підмножинах ознак, після чого агрегує їх прогнози шляхом голосування (для класифікації) або усереднення (для регресії) [36].

Процес побудови випадкового лісу включає два рівні випадкового розподілу. Кожне дерево навчається на окремій вибірці навчальних даних. Також при побудові кожного вузла дерева розглядається лише випадкова підмножина ознак замість всіх доступних характеристик. Алгоритм побудови моделей за цим методом наступний [36]:

1. Виберіть випадкові K точок даних з навчального набору даних.
2. Побудуйте дерева рішень, пов'язані з вибраними точками даних (підмножинами).
3. Виберіть число N для кількості дерев рішень, які ви хочете побудувати.
4. Повторіть перші два кроки.
5. Для нових точок даних знайдіть прогнози кожного дерева рішень і віднесіть нові точки даних до категорії, яка набрала більшість голосів.

У контексті прогнозування відтоку цей підхід може ефективно виявляти складні взаємодії між різними характеристиками співробітників. Наприклад, алгоритм може виявити, що поєднання певного рівня зарплати з конкретним типом проєктів та частотою зворотного зв'язку від керівника створює специфічний профіль ризику, який не очевидний при аналізі кожного фактора окремо.

Основними перевагами випадкового лісу є висока точність, стійкість до перенавчання та здатність працювати з великими наборами даних високої розмірності без попередньої обробки. Алгоритм також надає оцінки важливості ознак, що допомагає HR-фахівцям зрозуміти, які характеристики найбільше впливають на рішення працівників про звільнення. Алгоритм також менш чутливий до викидів порівняно з окремими деревами рішень завдяки усередненню прогнозів множини моделей.

Проте варто бути обережними, оскільки є висока вірогідність перенавчання при дуже великій кількості дерев. Алгоритм також може бути менш ефективним при роботі з дуже розрідженими даними або при наявності дуже сильних ознак, які домінують над іншими.

1.1.7. Градієнтний бустинг (Gradient boosting)

Градієнтний бустинг є ансамблевим методом машинного навчання, який послідовно поєднує велику кількість слабких моделей, які складаються зі зазвичай з дерев рішень, де кожна наступна модель намагається компенсувати помилки попередніх. Основна ідея методу полягає у поступовому зменшенні функції втрат за допомогою методу градієнтного спуску, що дозволяє будувати адаптивні моделі [37].

Дерева рішень з градієнтним підсиленням об'єднують кілька слабких учнів для створення одного сильного учня. Окремі дерева рішень в цьому випадку є слабкими учнями. Всі дерева з'єднуються послідовно, при цьому кожне дерево намагається зменшити помилку попереднього. Алгоритми, що швидко навчаються, часто важко піддаються навчанню, але є надзвичайно точними завдяки такому послідовному зв'язку. Моделі з повільною кривою навчання краще справляються зі статистичним навчанням [37].

Слабкі учні інтегруються таким чином, що кожен новий учень інтегрує залишки попереднього етапу в міру розвитку моделі. Фінальна модель об'єднує результати кожного етапу, що призводить до появи сильного учня. Залишки визначаються за допомогою функції втрат.

Серед переваг градієнтного бустингу варто виділити високу точність прогнозування, здатність працювати з різними типами даних, а також ефективне виявлення складних взаємозв’язків між змінними. Крім того, метод добре справляється з проблемою нерівномірного розподілу класів і дозволяє враховувати велику кількість ознак.

До недоліків методу відносять високу обчислювальну складність, чутливість до параметрів налаштування: кількість дерев, глибина, швидкість навчання, а також ризик перенавчання за надмірної кількості ітерацій. Крім того, результати моделі можуть бути менш інтерпретованими порівняно з класичними алгоритмами.

1.1.8. Нейронні мережі (Neural Networks)

Штучні нейронні мережі є потужним інструментом для моделювання складних нелінійних залежностей у даних. Вони складаються з шарів штучних нейронів, кожен з яких виконує операції зважування вхідних сигналів, застосування функцій активації та передачу результату на наступний рівень. Завдяки багатошаровій структурі нейронні мережі здатні формувати узагальнені представлення даних та виявляти приховані закономірності [38].

Нейронна мережа в машинному навчанні – це програмно запрограмована мережа штучних нейронів. Вона намагається імітувати людський мозок, маючи кілька шарів "нейронів", подібно до нейронів нашого мозку. Фото, відео, звук, текст та інші вхідні дані надходять до першого шару нейронів. Ці дані проходять через усі шари, причому вихід одного шару подається на наступний. Це особливо важливо для такої складної задачі, як обробка природної мови для машинного навчання [38].

Основними перевагами нейронних мереж є висока точність у завданнях класифікації та прогнозування, здатність до навчання на великих масивах даних, гнучкість у роботі з різними типами інформації, а також можливість моделювати складні багатовимірні залежності. Нейронні мережі добре пристосовані для завдань, де класичні алгоритми демонструють обмежену ефективність.

Водночас їхні недоліки включають значні обчислювальні витрати, потребу у великій кількості даних для навчання, складність налаштування гіперпараметрів (кількість шарів, нейронів, функції активації, швидкість навчання тощо). Окремою проблемою є низька інтерпретованість результатів, що ускладнює пояснення логіки прийняття рішень моделі.

1.1.9. Матриця помилок (Confusion Matrix)

Правильна оцінка якості моделей є критично важливою для успішного впровадження систем прогнозування відтоку працівників. Вибір відповідних метрик оцінки залежить від специфіки бізнес-задачі, балансу класів у даних та відносної важливості різних типів помилок [39].

Матриця помилок (Confusion Matrix) є основним інструментом для аналізу продуктивності моделей бінарної класифікації. Вона представляється у вигляді

таблиці, де кожен ряд відповідає фактичному класу, а кожний стовпець – прогнозованому класу. Зазвичай вважаються два класи: "позитивний" (наприклад, "працівник, який звільниться") та "негативний" (наприклад, " працівник, який залишається"). Однак, у випадку багатокласової класифікації, матриця помилок може мати більше рядків і стовпців, які відповідатимуть різним класам [39].

Під час бінарної класифікації матриця помилок використовує чотири основних значення:

- True Positives (TP): кількість працівників, які дійсно звільнилися і були правильно ідентифіковані моделлю;
- True Negatives (TN): кількість працівників, які залишилися в компанії і були правильно класифіковані;
- False Positives (FP): кількість працівників, які залишилися, але були помилково класифіковані як ті, що звільняться;
- False Negatives (FN): кількість працівників, які звільнилися, але не були виявлені моделлю.

За допомогою цих значень можна обчислити показники, що характеризують продуктивність моделі, такі як [39]:

1. F-міра (F1-score): Цей показник є гармонічним середнім між точністю і чутливістю і використовується для оцінки збалансованості моделі.

$$F1\text{-score} = 2 * (Precision * Recall) / (Precision + Recall) \quad (1.5)$$

2. Точність (Precision): Цей показник вимірює, наскільки точно модель класифікує позитивні приклади.

$$Precision = TP / (TP + FP) \quad (1.6)$$

3. Чутливість (Recall) або Сприйнятливість (Sensitivity): Цей показник вимірює, наскільки ефективно модель розпізнає позитивні приклади.

$$Recall = TP / (TP + FN) \quad (1.7)$$

4. Специфічність (Specificity): Цей показник вимірює, наскільки ефективно модель розпізнає негативні приклади.

$$Specificity = TN / (TN + FP) \quad (1.8)$$

5. Точність прогнозування (Accuracy): Цей показник вимірює загальну точність моделі і враховує як правильно класифіковані позитивні, так і негативні приклади.

$$Accuracy = (TP + TN) / (TP + TN + FP + FN) \quad (1.9)$$

6. Збалансована точність (Balanced Accuracy): Цей показник приділяє однакову вагу кожному класу, незалежно від його частки в сукупності спостережень. Він дозволяє оцінити загальну продуктивність моделі, враховуючи незбалансованість класів, коли кількість спостережень різних класів може суттєво відрізнятися.

$$Balanced\ Accuracy = (((TP/(TP+FN)+(TN/(TN+FP))) / 2 \quad (1.10)$$

$$Balanced\ Accuracy = (Sensitivity + Specificity) / 2 \quad (1.11)$$

Ці показники допомагають оцінити продуктивність моделі класифікації та зрозуміти, наскільки вона ефективно розпізнає кожен з класів. Зазвичай, вибір показника залежить від специфіки задачі та вимог до моделі. Наприклад, якщо важливо мінімізувати FP (False Positives) результати, то використовують точність (Precision). Якщо важливо мінімізувати FN (False Negative) результати, то використовують чутливість (Recall) [39].

Отже, машинне навчання представляє собою потужний інструмент для прогнозування відтоку працівників в IT-компаніях, який може значно підвищити ефективність стратегій утримання персоналу. Різноманітність доступних алгоритмів, від простих і інтерпретованих методів, таких як логістична регресія та дерева рішень, до складних ансамблевих підходів та глибинних нейронних мереж дозволяє обирати оптимальні рішення залежно від специфіки даних та бізнес-вимог.

Ключовими факторами успіху є якість підготовки даних, правильний вибір метрик оцінки з урахуванням незбалансованості класів, врахування часової динаміки HR-процесів та забезпечення інтерпретованості результатів. Особливої уваги потребують етичні аспекти використання алгоритмів у HR-процесах, включаючи питання справедливості та недискримінації.

Практичне впровадження систем прогнозування відтоку вимагає комплексного підходу, який включає технічну реалізацію, інтеграцію з існуючими системами,

навчання персоналу та постійний моніторинг ефективності. Успішна реалізація таких систем може принести значну економічну вигоду компаніям через зменшення витрат на найм, збереження інтелектуального капіталу та підвищення загальної ефективності управління персоналом.

Розвиток технологій машинного навчання та зростання обсягів доступних HR-даних створюють нові можливості для підвищення точності та корисності систем прогнозування відтоку працівників. Майбутні дослідження можуть фокусуватися на розробці більш складних моделей, що враховують соціальні мережі всередині організації, аналізі неструктурованих даних та створенні адаптивних систем, здатних автоматично пристосовуватися до змін у бізнес-середовищі.

1.4. Огляд готових рішень на ринку

Останніми роками прогнозування плинності кадрів перетворилося на окремий напрям HR-аналітики, і на ринку з’явилася низка готових програмних рішень, спрямованих на підтримку компаній у зменшенні рівня звільнень та підвищенні стабільності кадрового складу. Ці рішення поєднують методи машинного навчання, аналіз великих масивів даних та інтеграцію з HRM-системами, що дозволяє компаніям оперативно виявляти ризики й ухвалювати управлінські рішення.

Серед найбільш відомих можна виокремити SAP SuccessFactors, Workday та Oracle HCM Cloud. Усі ці системи містять модулі прогнозної аналітики, що дозволяють ідентифікувати співробітників із підвищеним ризиком звільнення. Наприклад, SAP SuccessFactors застосовує алгоритми машинного навчання для аналізу даних про продуктивність, історію роботи, рівень компенсації та результати опитувань, формуючи інтегральний індекс ризику відтоку. Workday використовує хмарні технології для збору й обробки інформації про життєвий цикл працівника та надає менеджерам інтерпретовані прогнози з рекомендаціями щодо утримання. Oracle HCM Cloud інтегрує аналітичні панелі з візуалізацією ключових факторів ризику, забезпечуючи HR-фахівців прозорими інструментами прийняття рішень.

Окрему категорію становлять спеціалізовані платформи з аналітики персоналу, серед яких Visier People, UKG Pro та ADP DataCloud. Вони надають поглиблену

аналітику щодо відтоку кадрів, використовуючи моделі прогнозування та автоматичні дашборди для виявлення тенденцій у плинності залежно від підрозділів, посад чи демографічних характеристик. Завдяки інтеграції з системами обліку робочого часу та фінансовими модулями ці рішення дозволяють поєднувати прогнозування ризику звільнення із розрахунком економічних наслідків.

Окрім корпоративних HRM-платформ, на ринку існують і вузькоспеціалізовані рішення, орієнтовані саме на завдання прогнозування відтоку. Прикладами є сервіси PredictiveHR та Eightfold.ai, які активно застосовують методи глибинного навчання. Вони пропонують компаніям інструменти для оцінки ймовірності звільнення працівників, аналізу їхнього професійного розвитку та навіть формування персоналізованих кар’єрних рекомендацій.

Узагальнюючи, можна відзначити, що готові рішення на ринку значно відрізняються за функціональністю, рівнем інтеграції та вартістю. Для великих міжнародних компаній найчастіше використовуються комплексні HRM-платформи SAP, Workday, Oracle, тоді як середні й малі підприємства можуть обирати більш гнучкі й вузькоспеціалізовані сервіси PredictiveHR, Eightfold.ai. Незважаючи на різницю в підходах, усі ці рішення демонструють високий попит та підтверджують актуальність завдання прогнозування плинності кадрів у сучасних умовах.

ВИСНОВКИ ДО РОЗДІЛУ 1

У першому розділі було здійснено теоретичний аналіз сучасних засад управління утриманням працівників IT-компаній, що дозволило сформувати цілісне уявлення про предмет дослідження та визначити підходи, релевантні для подальшої розробки моделі прогнозування відтоку персоналу.

По-перше, розгляд сучасних методологій управління проєктами у сфері Data Science засвідчив, що традиційні підходи подібні до Waterfall, не забезпечують необхідної гнучкості та адаптивності в умовах високої невизначеності та змінності аналітичних завдань. Натомість застосування гнучких методологій, об’єднаних у парадигму Agile (Scrum, Kanban, Data-Driven Scrum, Agile Data Science), а також таких процесних моделей, як CRISP-DM, KDD, SEMMA виявляється більш ефективним. Ці підходи забезпечують ітеративність, швидке реагування на зміни та інтеграцію бізнес- і технічних вимог, що є визначальним для успішної реалізації проєктів із розробки моделей машинного навчання.

По-друге, дослідження особливостей та підходів до утримання працівників в IT-сфері виявило низку чинників, що зумовлюють високу плинність кадрів. Серед ключових причин звільнень і незадоволеності працівників виокремлюються відсутність можливостей професійного розвитку, робота із застарілими технологіями, недостатня конкурентоспроможність компенсації, високе навантаження, вигорання та неефективна організаційна культура. Успішні стратегії утримання передбачають формування бренду роботодавця, створення сприятливого робочого середовища, запровадження індивідуалізованих систем мотивації, підтримку балансу між роботою та особистим життям, а також активне використання HR-аналітики для своєчасного виявлення ризикових груп працівників.

По-третє, огляд методів машинного навчання для прогнозування відтоку працівників довів, що такі алгоритми, як логістична регресія, дерева рішень, Random Forest, XGBoost, нейронні мережі та ансамблеві моделі, забезпечують високу точність прогнозів за умови якісної підготовки даних та коректного відбору ознак.

Використання цих методів дозволяє не лише передбачати можливе звільнення, але й виявляти ключові фактори, що впливають на лояльність і задоволеність персоналу. Водночас особливого значення набуває інтеграція кількісних результатів із соціально-психологічними аспектами, які традиційно важко формалізувати.

Окремо було розглянуто готові програмні рішення на ринку, що пропонують інструменти для прогнозування плинності кадрів, такі як SAP SuccessFactors, Workday, Oracle HCM Cloud, Visier, PredictiveHR, Eightfold.ai. Вони демонструють практичну значущість таких моделей та їхню затребуваність у бізнес-середовищі, проте відрізняються за вартістю, гнучкістю інтеграції та можливістю адаптації до специфіки конкретної компанії. Це підкреслює доцільність створення власних моделей для IT-компаній, які здатні враховувати індивідуальні потреби, корпоративну культуру та особливості внутрішніх процесів.

Отже, результати проведеного теоретичного аналізу підтвердили, що управління утриманням працівників в IT-компаніях є багатокомпонентним завданням, яке вимагає інтеграції сучасних підходів проєктного менеджменту, HR-стратегій та інструментів машинного навчання. Для забезпечення стійкості кадрового потенціалу та підвищення конкурентоспроможності IT-компаній необхідно поєднувати індивідуалізовані методи мотивації та розвитку персоналу з аналітичними інструментами прогнозування ризиків звільнення. Таким чином, результати першого розділу створюють основу для подальшої розробки та тестування моделей прогнозування утримання працівників.

РОЗДІЛ 2. РОЗРОБКА ТА ТЕСТУВАННЯ МОДЕЛЕЙ ПРОГНОЗУВАННЯ УТРИМАННЯ ПРАЦІВНИКІВ

2.1. Формування та підготовка даних для аналізу

Плинність кадрів у сфері інформаційних технологій є однією з найгостріших проблем сучасного бізнесу. Висока вартість підбору, навчання та адаптації нових співробітників, а також втрата накопичених знань і досвіду роблять завдання утримання персоналу критично важливим для успіху. За статистикою, вартість заміщення одного IT-спеціаліста може становити понад 50% його річної заробітної плати, що підкреслює економічну доцільність інвестицій у системи прогнозування та запобігання плинності кадрів [5].

Для проведення дослідження використовуватиметься набір даних від компанії IBM, яка надала є одним із найбільш поширених і якісних наборів даних для аналізу плинності кадрів. Цей набір даних містить інформацію про 1470 співробітників компанії та включає 35 різних характеристик, що описують демографічні, професійні, фінансові та особистісні аспекти працівників [40].

Датасет був створений компанією IBM з метою демонстрації можливостей аналітики в управлінні людськими ресурсами та надання дослідникам якісного інструменту для розробки моделей прогнозування плинності кадрів. Ці дані максимально наближені до реальних HR-даних і враховують основні фактори, що впливають на рішення співробітників про залишення компанії. Датасет описує наступні характеристики:

- Демографічні ознаки:
    - Age – вік співробітника.
    - Gender – стать (Male/Female).
    - MaritalStatus – сімейний стан (Single/Married/Divorced).
    - DistanceFromHome – відстань від дому до роботи у милях.
- Освітні ознаки:
    - Education – рівень освіти.

  - EducationField – галузь освіти (Life Sciences, Medical, Marketing, Technical Degree, Human Resources, Other).
- Професійні ознаки:
  - Department – відділ.
  - JobRole – назва посади.
  - JobLevel – рівень посади.
  - JobInvolvement – рівень залученості у роботу.
  - JobSatisfaction – рівень задоволеності роботою.
  - PerformanceRating – оцінка ефективності роботи.
- Фінансові ознаки:
  - MonthlyIncome, MonthlyRate, DailyRate, HourlyRate – числові значення зарплати у доларах США.
  - PercentSalaryHike – відсоток останнього підвищення зарплати.
  - StockOptionLevel – рівень опціонів, якими володіє працівник.
- Досвід:
  - TotalWorkingYears, YearsAtCompany, YearsInCurrentRole, YearsSinceLastPromotion, YearsWithCurrManager – числові змінні у роках.
  - NumCompaniesWorked – кількість компаній, у яких раніше працював співробітник.
- Робочі умови:
  - BusinessTravel – частота відряджень.
  - OverTime – факт роботи понаднормово (Yes/No).
  - WorkLifeBalance – оцінка рівня балансу роботи та особистого життя.
- Цільова ознака прогнозування: Attrition – факт звільнення (Yes/No).
- Інші ознаки:
  - EnvironmentSatisfaction – задоволеність робочим середовищем.
  - RelationshipSatisfaction – задоволеність взаєминами у колективі.
  - TrainingTimesLastYear – кількість тренінгів за рік.
  - EmployeeNumber – унікальний ідентифікатор.

Отже, цільовою (залежною) змінною є Attrition, яка вказує, чи звільнився працівник. Ми використаємо наявні ознаки для створення моделі, яка допоможе нам передбачити відтік працівників компанії. Таким чином, ми матимемо згоду приймати належні заходи зі зменшення % відтоку та зменшення загальних втрат на залучення нових працівників.

2.2. Дослідницький аналіз даних (EDA)

Дослідницький аналіз даних (Exploratory Data Analysis) є важливим етапом у процесі обробки та аналізу. Він передбачає первинне вивчення даних з метою виявлення основних закономірностей, аномалій та взаємозв'язків між змінними, ще до застосування складних методів машинного навчання. Основною метою EDA є глибоке розуміння структури даних, перевірка гіпотез, підготовка набору даних для подальшої обробки та забезпечення надійності результатів. EDA дозволяє швидко оцінити якість даних, виявити пропуски або некоректні значення та сформувати стратегію подальшої роботи з ними. Для проведення аналізу і побудови моделей в подальшому використовуватиметься мова програмування Python.

2.2.1. Первинний огляд даних

Початковий етап дослідження передбачає імпорт необхідних програмних бібліотек для статистичного аналізу та побудови моделей машинного навчання (Додаток А, Лістинг 1).

Наступним кроком є завантаження набору даних від IBM, що містить відомості про плинність кадрів. Для початкового ознайомлення з даними здійснено огляд окремих стовпців перших п'яти записів (Додаток А, Лістинг 2).

```
Розмір датасету: (1470, 35)

   Age Attrition     BusinessTravel  DailyRate              Department  \
0   41       Yes      Travel_Rarely       1102                   Sales
1   49        No  Travel_Frequently        279  Research & Development
2   37       Yes      Travel_Rarely       1373  Research & Development
3   33        No  Travel_Frequently       1392  Research & Development
4   27        No      Travel_Rarely        591  Research & Development

   DistanceFromHome  Education EducationField  EmployeeCount  EmployeeNumber  \
0                 1          2  Life Sciences              1               1
1                 8          1  Life Sciences              1               2
2                 2          2          Other              1               4
3                 3          4  Life Sciences              1               5
4                 2          1        Medical              1               7

   ...  RelationshipSatisfaction StandardHours  StockOptionLevel  \
0  ...                         1            80                 0
1  ...                         4            80                 1
2  ...                         2            80                 0
3  ...                         3            80                 0
4  ...                         4            80                 1

   TotalWorkingYears  TrainingTimesLastYear WorkLifeBalance  YearsAtCompany  \
0                  8                      0               1               6
1                 10                      3               3              10
2                  7                      3               3               0
3                  8                      3               3               8
4                  6                      3               3               2

  YearsInCurrentRole  YearsSinceLastPromotion  YearsWithCurrManager
0                  4                        0                     5
1                  7                        1                     7
2                  0                        0                     0
3                  7                        3                     0
4                  2                        2                     2
```

Рис. 2.1. Результати перегляду перших записів у наборі даних

**Джерело: створено автором [Додаток А, Лістинг 2]*

Детальне вивчення структури даних включає аналіз статистичних характеристик розподілу, пропусків, а також основних описових статистик для кількісних показників: середніх арифметичних, максимальних та мінімальних значень, медіан (Додаток А, Лістинг 3).

	Age	DailyRate	DistanceFromHome	Education	EmployeeCount
count	1470.000000	1470.000000	1470.000000	1470.000000	1470.0
mean	36.923810	802.485714	9.192517	2.912925	1.0
std	9.135373	403.509100	8.106864	1.024165	0.0
min	18.000000	102.000000	1.000000	1.000000	1.0
25%	30.000000	465.000000	2.000000	2.000000	1.0
50%	36.000000	802.000000	7.000000	3.000000	1.0
75%	43.000000	1157.000000	14.000000	4.000000	1.0
max	60.000000	1499.000000	29.000000	5.000000	1.0

	EmployeeNumber	EnvironmentSatisfaction	HourlyRate	JobInvolvement	\
count	1470.000000	1470.000000	1470.000000	1470.000000	
mean	1024.865306	2.721769	65.891156	2.729932	
std	602.024335	1.093082	20.329428	0.711561	
min	1.000000	1.000000	30.000000	1.000000	
25%	491.250000	2.000000	48.000000	2.000000	
50%	1020.500000	3.000000	66.000000	3.000000	
75%	1555.750000	4.000000	83.750000	3.000000	
max	2068.000000	4.000000	100.000000	4.000000	

	JobLevel	...	RelationshipSatisfaction	StandardHours	\
count	1470.000000	...	1470.000000	1470.0	
mean	2.063946	...	2.712245	80.0	
std	1.106940	...	1.081209	0.0	
min	1.000000	...	1.000000	80.0	
25%	1.000000	...	2.000000	80.0	
50%	2.000000	...	3.000000	80.0	
75%	3.000000	...	4.000000	80.0	
max	5.000000	...	4.000000	80.0	

	StockOptionLevel	TotalWorkingYears	TrainingTimesLastYear	\
count	1470.000000	1470.000000	1470.000000	
mean	0.793878	11.279592	2.799320	
std	0.852077	7.780782	1.289271	
min	0.000000	0.000000	0.000000	
25%	0.000000	6.000000	2.000000	
50%	1.000000	10.000000	3.000000	
75%	1.000000	15.000000	3.000000	
max	3.000000	40.000000	6.000000	

	WorkLifeBalance	YearsAtCompany	YearsInCurrentRole	\
count	1470.000000	1470.000000	1470.000000	
mean	2.761224	7.008163	4.229252	
std	0.706476	6.126525	3.623137	
min	1.000000	0.000000	0.000000	
25%	2.000000	3.000000	2.000000	
50%	3.000000	5.000000	3.000000	
75%	3.000000	9.000000	7.000000	
max	4.000000	40.000000	18.000000	

	YearsSinceLastPromotion	YearsWithCurrManager
count	1470.000000	1470.000000
mean	2.187755	4.123129
std	3.222430	3.568136
min	0.000000	0.000000
25%	0.000000	2.000000
50%	1.000000	3.000000
75%	3.000000	7.000000
max	15.000000	17.000000

Рис. 2.2. Статистичні характеристики даних

**Джерело: створено автором [Додаток А, Лістинг 3]*

Дані не містять пропусків, проте після детальної перевірки було знайдено, що дані містять константі значення у трьох ознак: EmployeeCount, StandardHours, Over18. Також наявний унікальний ідентифікатор кожного працівника EmployeeNumber. Такі поля не беруть участі у формуванні закономірностей, а отже, будуть вилучені на початковому етапі. Загалом, якість вихідних даних можна оцінити як високу, що

дозволяє зосередитися на більш глибокому аналізі без значних витрат часу на їх очищення (Додаток А, Лістинг 4).

2.2.2. Аналіз розподілу цільової змінної

Для визначення розподілу класів у датасеті проведено аналіз частотності значень змінної Attrition, що дозволяє виявити потенційний дисбаланс між кількістю працівників, які залишились в компанії, та тими, хто звільнився (Додаток А, Лістинг 4). Результати показують значну нерівномірність розподілу: з 1470 спостережень лише 237 записів (16,1%) відповідають випадкам звільнення, тоді як 1233 записи (83,9%) стосуються працівників, що продовжують роботу в організації (Див. Рис. 2.3). Така диспропорція свідчить про наявність істотного дисбалансу класів, що може негативно впливати на якість навчання моделей машинного навчання та призводити до упередженості у бік переважаючого класу. Це обумовлює необхідність застосування спеціальних методів роботи з незбалансованими даними на етапі побудови прогностичних моделей.

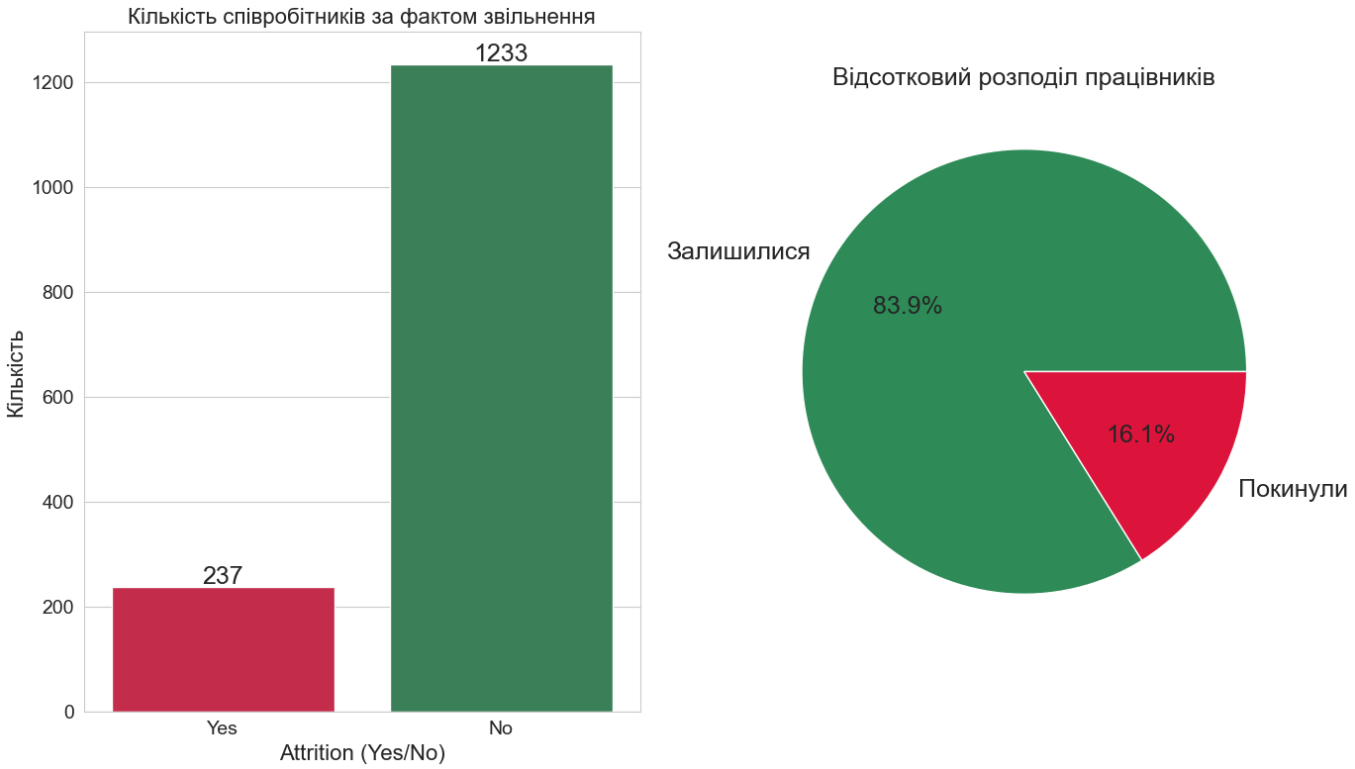

Рис. 2.3. Графіки розподілу цільової змінної

**Джерело: створено автором [Додаток А, Лістинг 4]*

2.2.3. Аналіз демографічних характеристик

У межах цього підпункту буде здійснено аналіз демографічних характеристик працівників, які можуть впливати на їх рішення залишитися в компанії чи звільнитися. Зокрема, досліджено такі змінні, як вік, стать, сімейний стан та відстань від місця проживання до роботи. Для побудови графіків використовуватиметься пакет Matplotlib та Seaborn.

Аналіз демографічних характеристик показав, що середній вік співробітників, які залишилися працювати в компанії, становить близько 37,7 років, тоді як ті, хто звільнився, мають середній вік 33,6 років. Це може свідчити про те, що молодші працівники більш схильні до зміни місця роботи, ймовірно, через більшу мобільність та прагнення швидшого кар'єрного зростання (Див. Рис. 2.4).

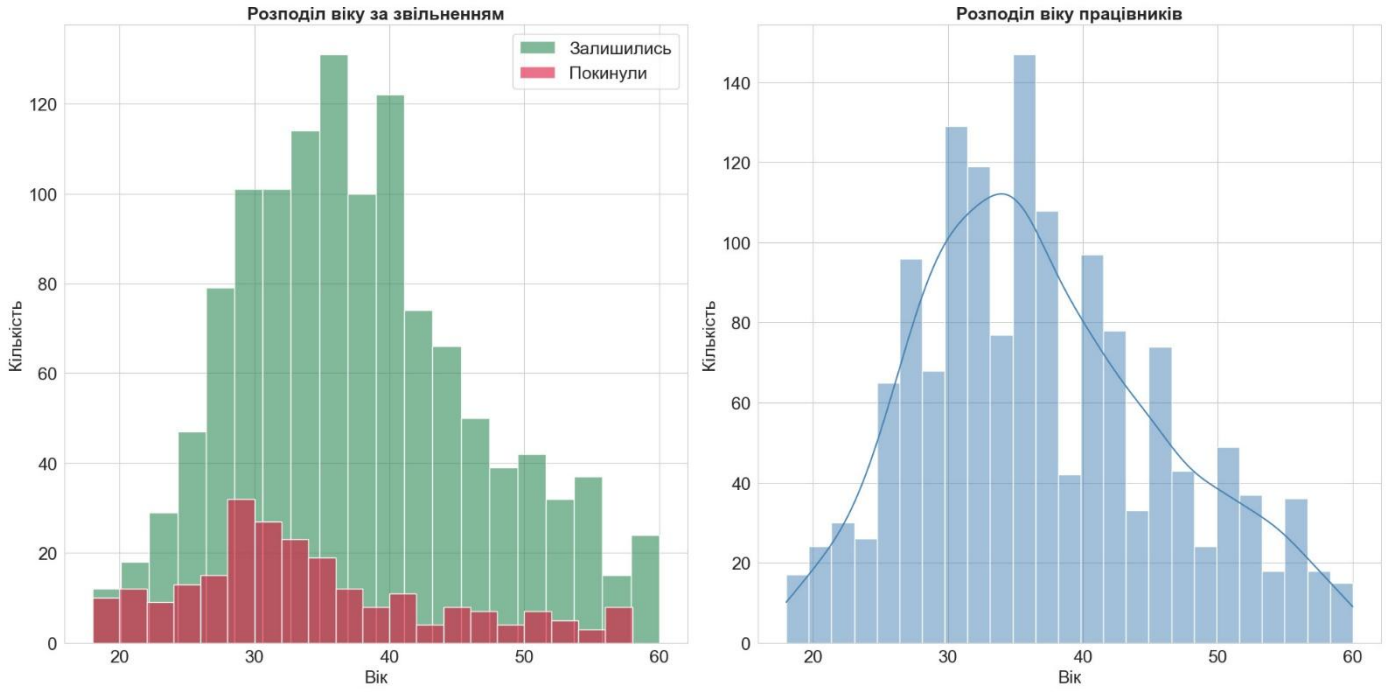

Рис. 2.4. Графіки розподілу за віком

**Джерело: створено автором [Додаток А, Лістинг 6, 7]*

Гендерний розподіл вказує на відносно рівні показники плинності серед чоловіків та жінок, хоча незначна перевага у рівні звільнень спостерігається серед чоловіків. Це може бути пов'язано як із культурними, так і зі структурними особливостями роботи в компанії (Див. Рис. 2.5).

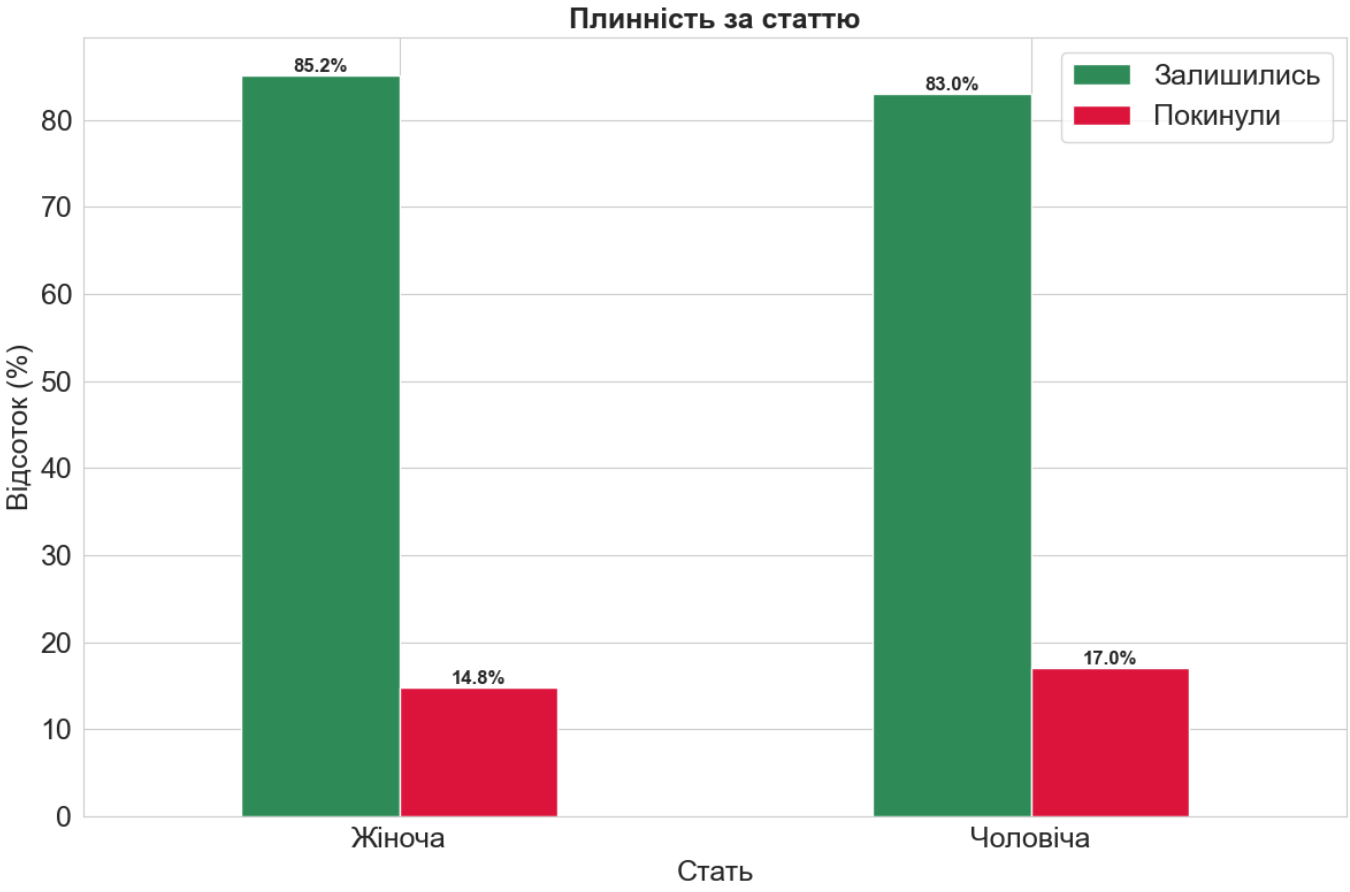

Рис. 2.5. Графік розподілу відтоку за статтю

**Джерело: створено автором [Додаток А, Лістинг 9]*

Сімейний стан виявився важливим фактором: найвищий рівень плинності характерний для працівників, які не перебувають у шлюбі, що може бути наслідком їхньої більшої схильності до змін і меншої прив'язаності до стабільності. У той же час, одружені співробітники демонструють значно нижчий рівень звільнень, що підтверджує гіпотезу про роль сімейних зобов'язань у збереженні робочого місця (Див. Рис. 2.6).

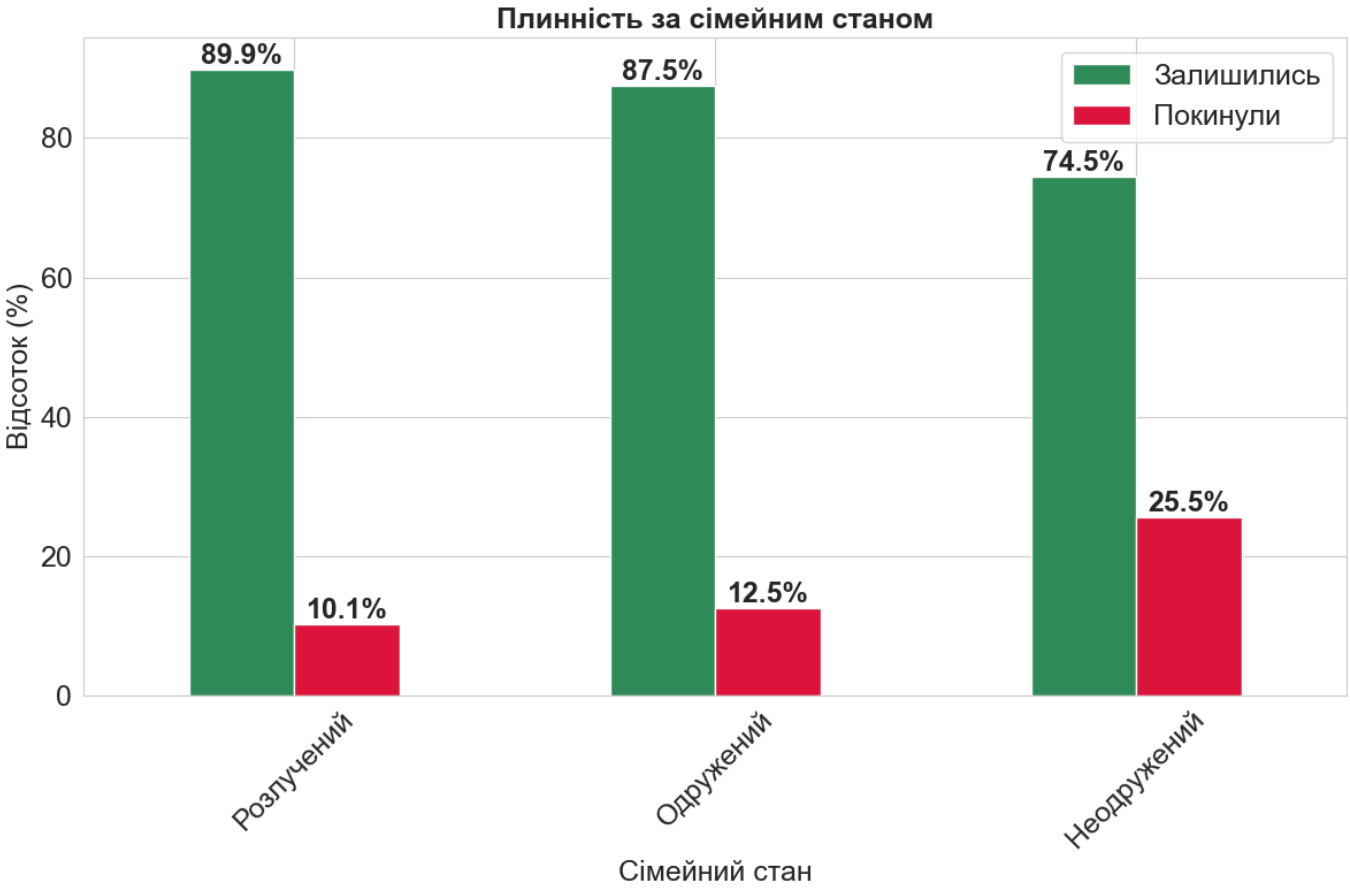

Рис. 2.6. Графік відтоку за сімейним станом

**Джерело: створено автором [Додаток А, Лістинг 10]*

Щодо відстані від дому до роботи, то співробітники, які проживають далі, демонструють вищу ймовірність звільнення. Це може бути зумовлено додатковими витратами часу та ресурсів на дорогу, що зн ижує загальну задоволеність умовами праці. Таким чином, демографічні характеристики мають помітний вплив на плинність кадрів і повинні враховуватися при побудові прогнозних моделей (Див. Рис. 2.7).

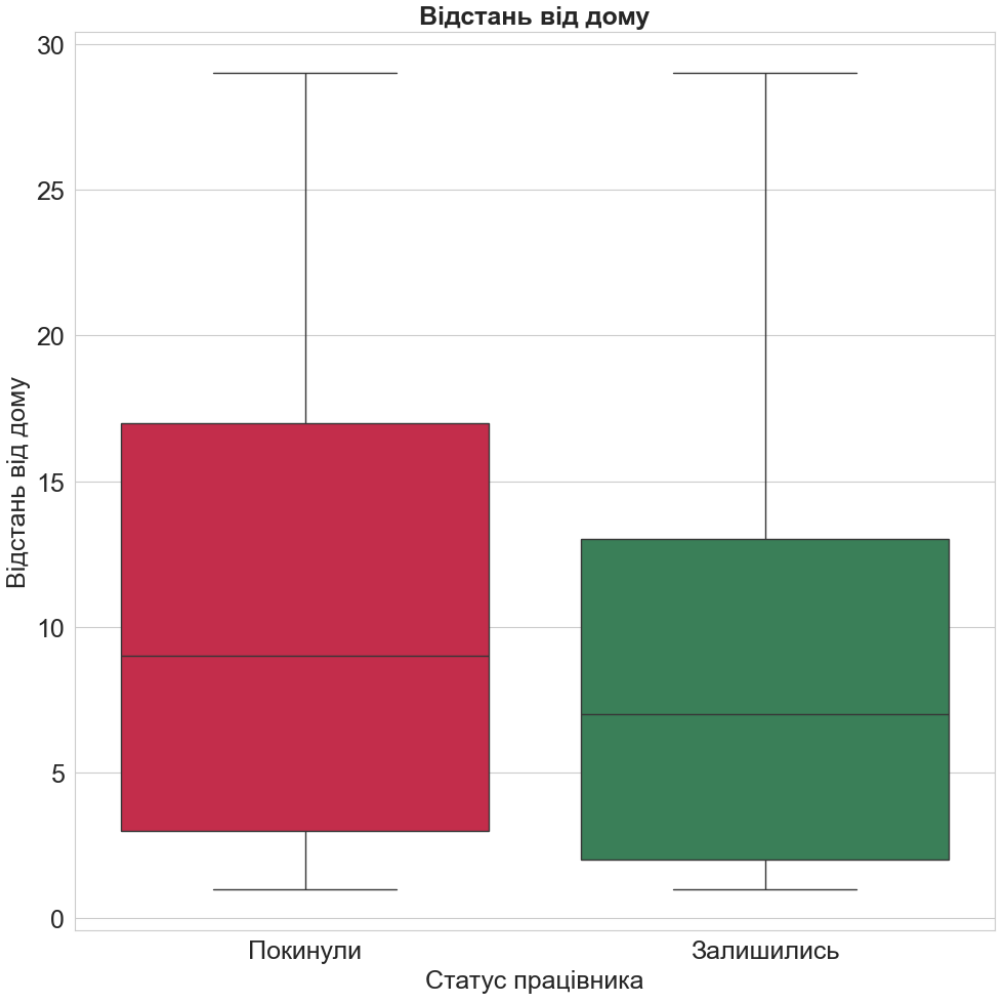

Рис. 2.7. Графік розподілу відстані від дому за відтоком

**Джерело: створено автором [Додаток А, Лістинг 11]*

2.2.4. Аналіз професійних характеристик

Професійні характеристики співробітників є ключовими чинниками, що можуть суттєво впливати на їхнє рішення залишитися в компанії або змінити місце роботи. До таких характеристик належать відділ, посада, рівень зайнятості та залученості, показники задоволеності працівників, оцінка їхньої ефективності, а також баланс між роботою і особистим життям. Дослідження цих змінних дозволяє виявити закономірності, притаманні різним групам працівників, та виявити потенційні проблемні зони в організації.

Найбільша частка звільнень спостерігається у відділі продажів, що може бути зумовлено високим рівнем конкуренції, стресовими умовами роботи та більшою кількістю альтернативних пропозицій на ринку праці. У відділі розробки рівень

плинності помітно нижчий, що свідчить про стабільніші умови праці та менший вплив зовнішніх факторів (Див. Рис. 2.8).

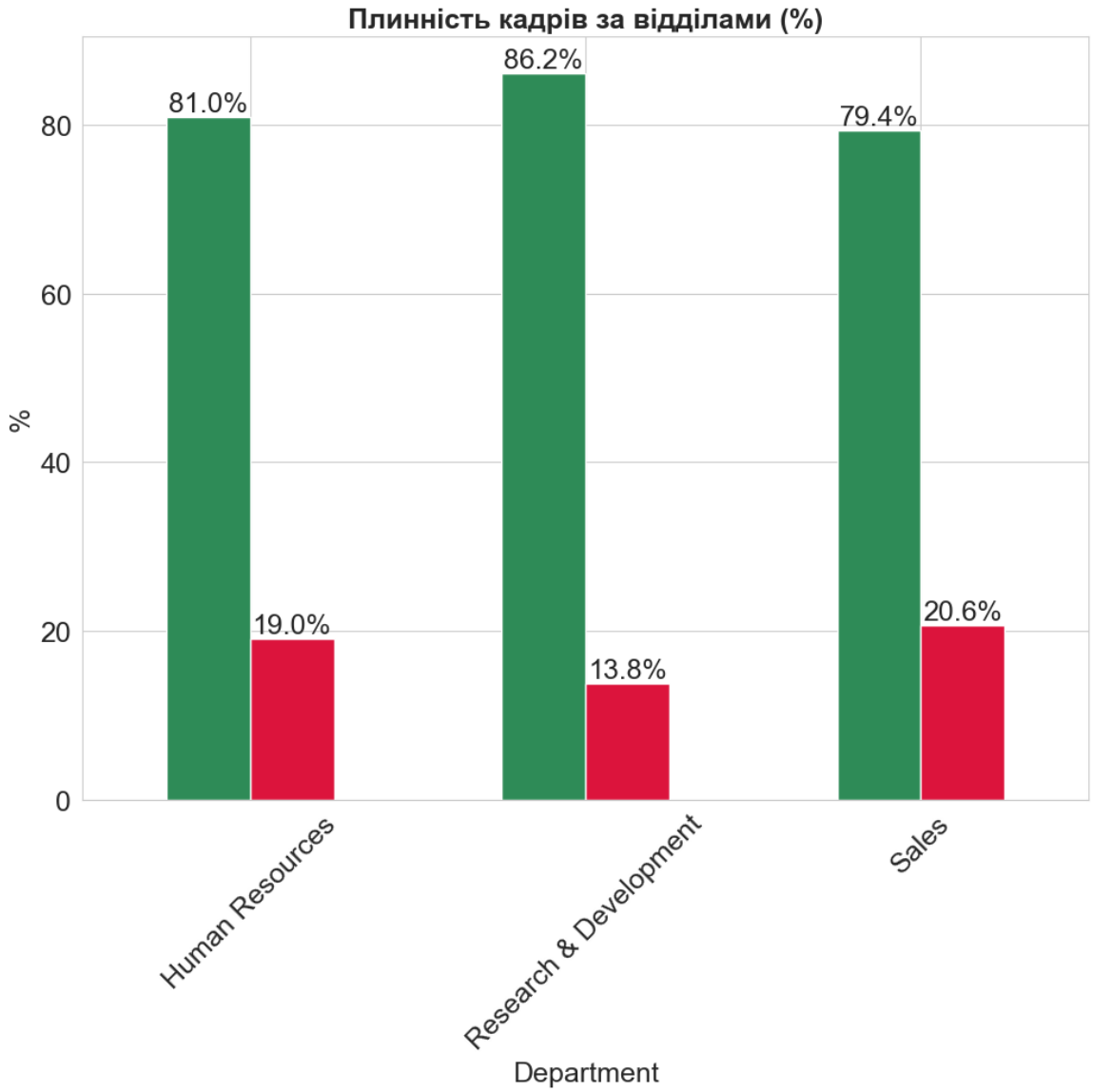

Рис. 2.8. Графік залежності відтоку за відділами

**Джерело: створено автором [Додаток А, Лістинг 12]*

Рівень посади також відіграє суттєву роль: співробітники з початковими позиціями демонструють найвищий рівень звільнень, тоді як для вищих посад цей показник знижується. Це може пояснюватися тим, що працівники з більшим досвідом та вищими посадами, як правило, мають кращі умови праці, вищий рівень задоволеності та фінансової компенсації (Див. Рис. 2.9).

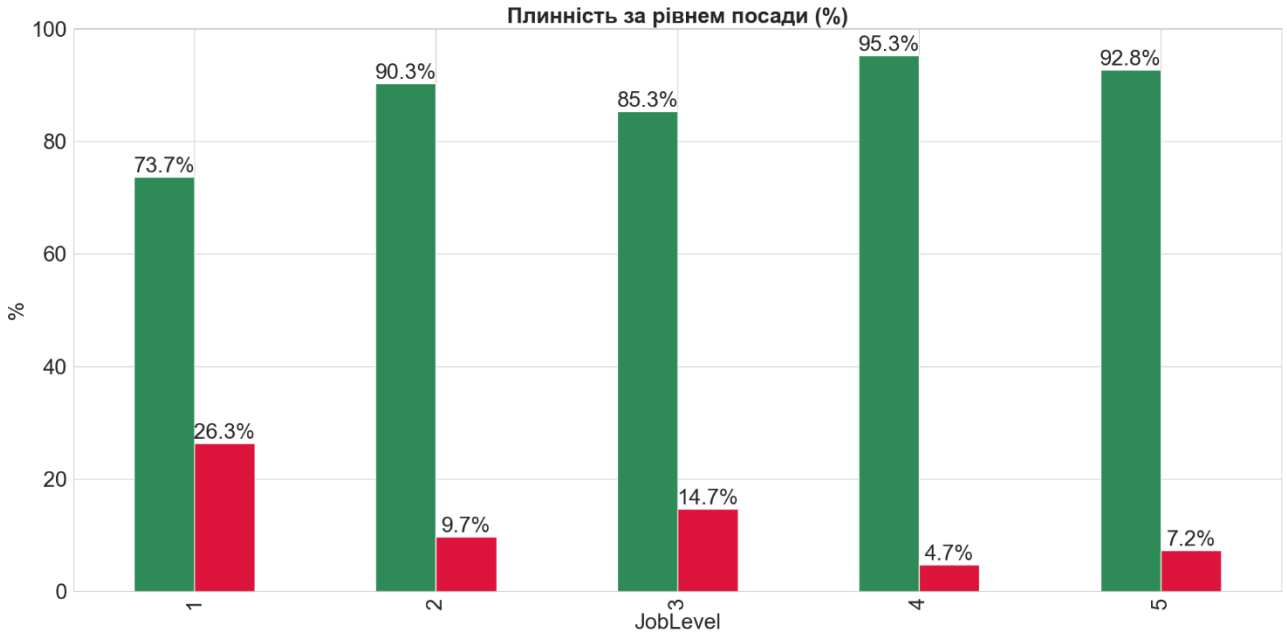

Рис. 2.9. Графік залежності відтоку за рівнем посади

**Джерело: створено автором [Додаток А, Лістинг 13]*

Аналіз показників задоволеності роботою та залученості у робочі процеси демонструє пряму залежність між цими факторами та ймовірністю залишення компанії. Працівники з низькими оцінками задоволеності та залученості частіше залишають компанію, тоді як високі значення цих показників асоціюються з вищою стабільністю працівників (Див. Рис. 2.10 та Рис. 2.11).

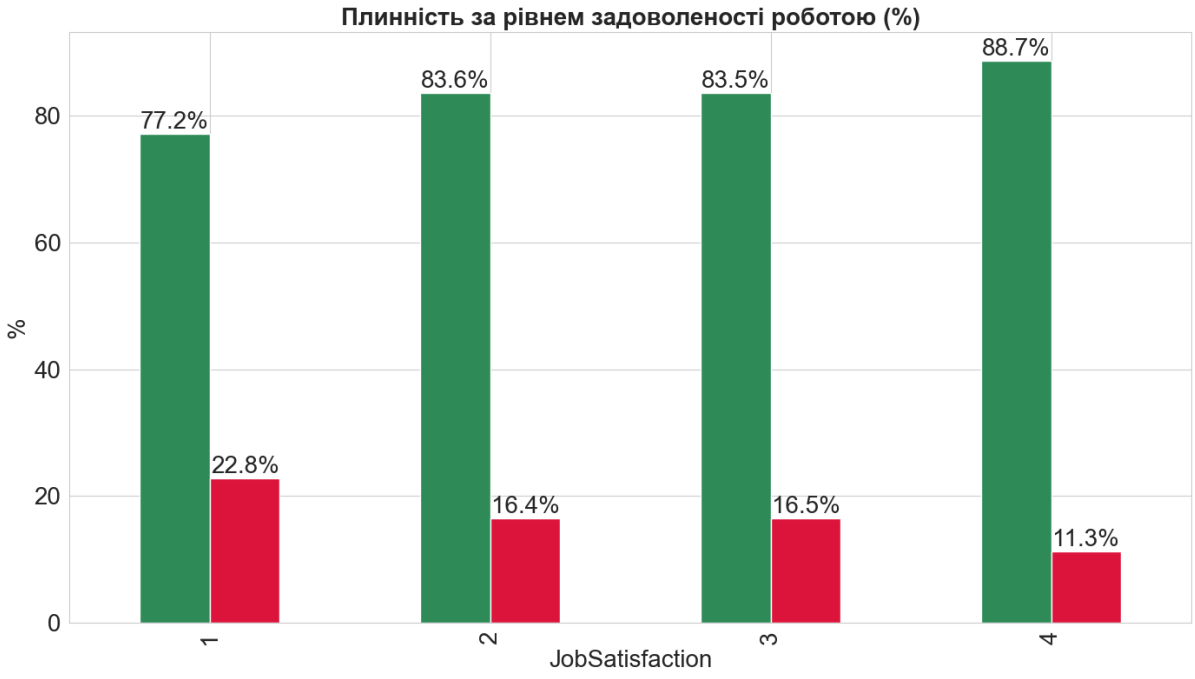

Рис. 2.10. Графік залежності відтоку за рівнем задоволеності роботою

**Джерело: створено автором [Додаток А, Лістинг 14]*

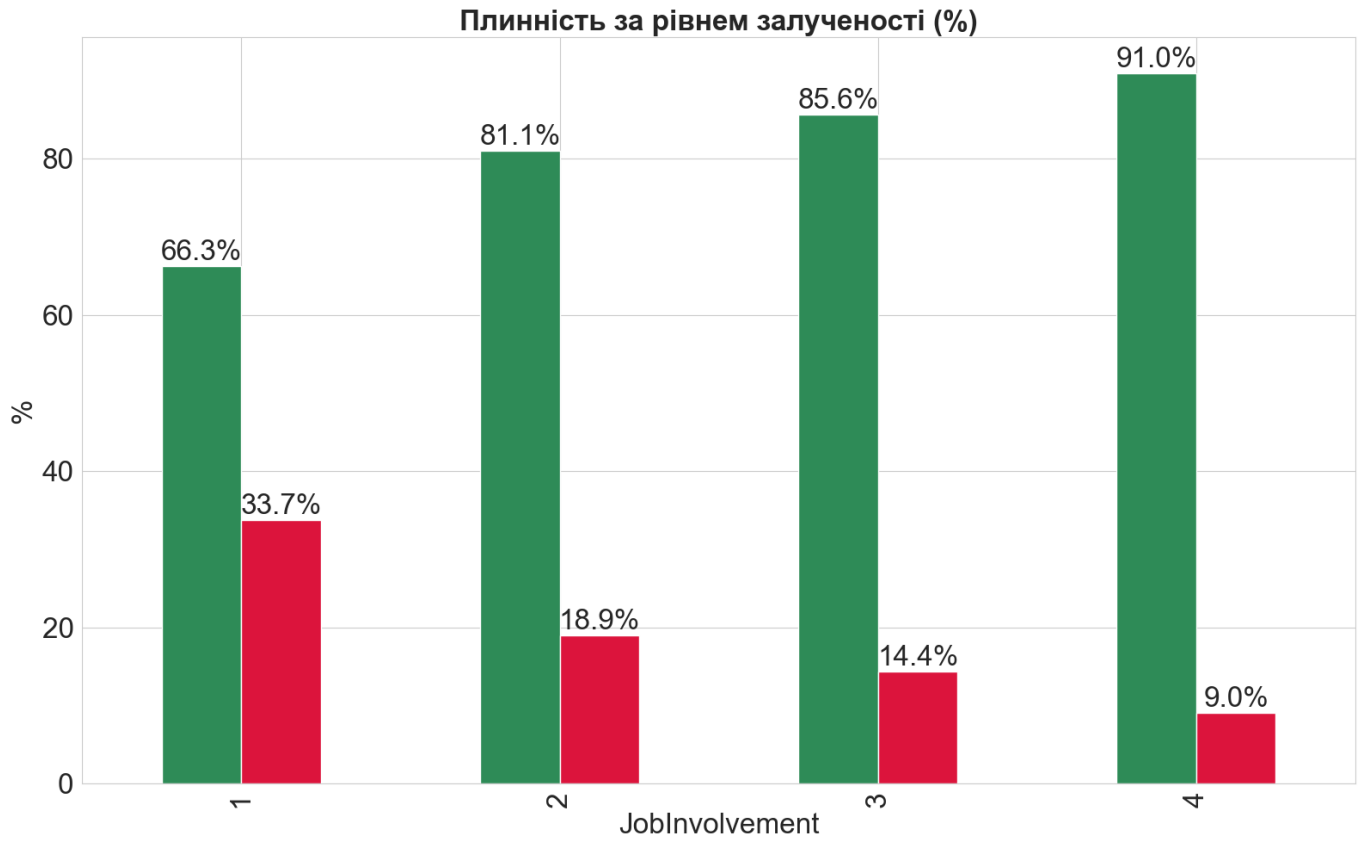

Рис. 2.11. Графік залежності відтоку за рівнем залученості у роботу

**Джерело: створено автором [Додаток А, Лістинг 15]*

Особливо значущим виявився фактор понаднормових годин. Співробітники, які працюють понаднормово, звільняються значно частіше, що може бути наслідком перевтоми та зниження мотивації. Також помітно, що часті відрядження збільшують імовірність звільнення, тоді як відсутність відряджень або їх рідкість асоціюються з більшою стабільністю кадрів (Див. Рис. 2.12 та Рис. 2.13).

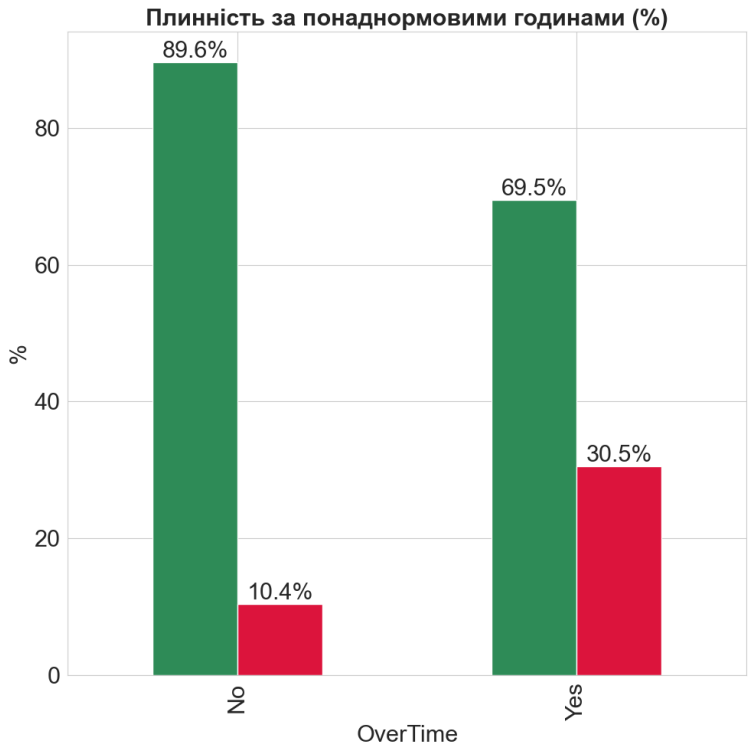

Рис. 2.12. Графік залежності відтоку від за наявністю понаднормових годин роботи

**Джерело: створено автором [Додаток А, Лістинг 18]*

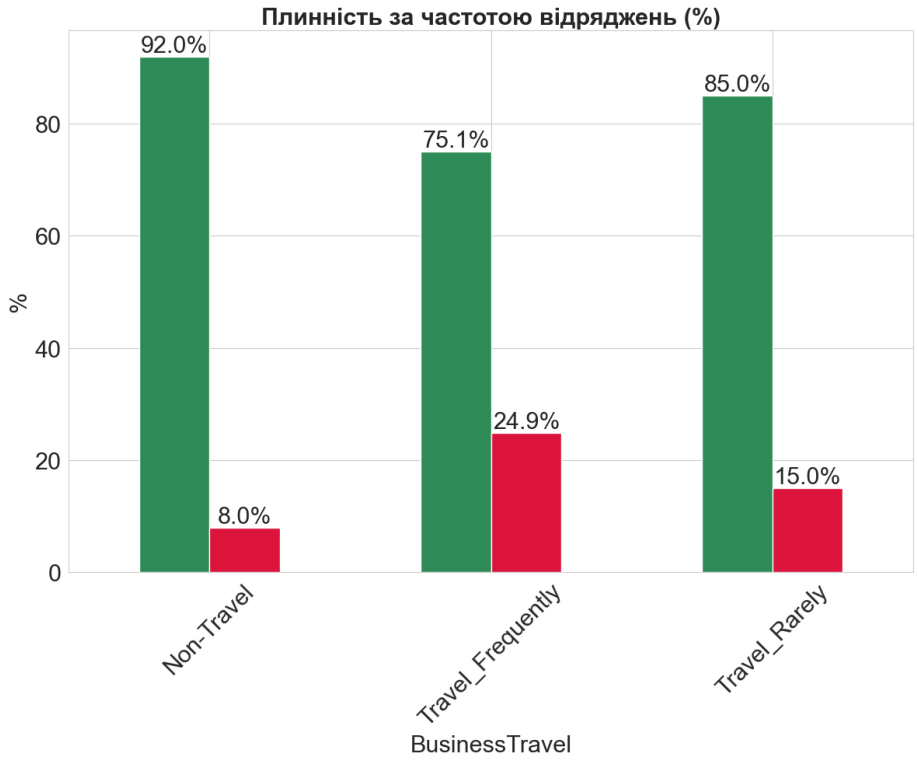

Рис. 2.13. Графік залежності відтоку за частотою відряджень

**Джерело: створено автором [Додаток А, Лістинг 19]*

Отримані результати підтверджують, що професійні характеристики є критично важливими для прогнозування плинності кадрів, а також вказують на

напрями, у яких HR-відділ може впливати на ситуацію, зокрема, шляхом покращення умов праці, оптимізації робочого навантаження та підвищення рівня задоволеності співробітників.

2.2.5. Аналіз досвіду та стажу роботи

Досвід та тривалість роботи співробітників у компанії також часто впливає на плинність кадрів. З одного боку, накопичений досвід і тривалий стаж можуть сприяти підвищенню лояльності та адаптації до корпоративної культури, з іншого – тривале перебування на одній посаді без кар'єрного зростання може призводити до зниження мотивації та збільшення ймовірності звільнення.

Аналіз показників, пов'язаних із досвідом і стажем роботи, виявив декілька закономірностей, які мають суттєве значення для прогнозування плинності кадрів. Загальний трудовий стаж співробітників, що залишили компанію, у середньому є нижчим, ніж у тих, хто її не покинув. Це може свідчити про те, що менш досвідчені працівники більш мобільні та схильні до пошуку нових можливостей (Див. Рис. 2.14).

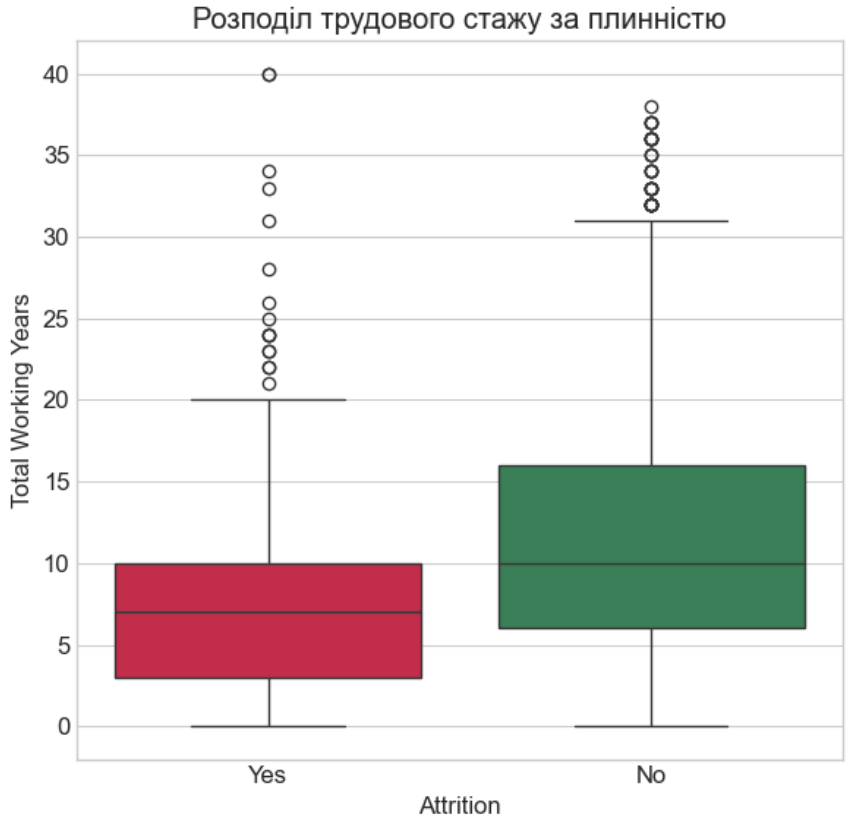

Рис. 2.14. Графік розподілу трудового стажу за плинністю

**Джерело: створено автором [Додаток А, Лістинг 20]*

Подібна тенденція спостерігається і для показника кількості років у компанії. Працівники, що пропрацювали менше часу, мають значно вищий рівень звільнень, що може бути пов’язано з недостатньою адаптацією до корпоративної культури та відсутністю сформованої лояльності (Див. Рис. 2.15). Водночас, серед тих, хто працює в компанії багато років, також зустрічаються випадки звільнень. Аналіз кількості років на поточній посаді показав, що співробітники, які тривалий час не змінювали роль, трохи частіше залишають компанію, особливо якщо відсутні можливості для підвищення (Див. Рис. 2.16).

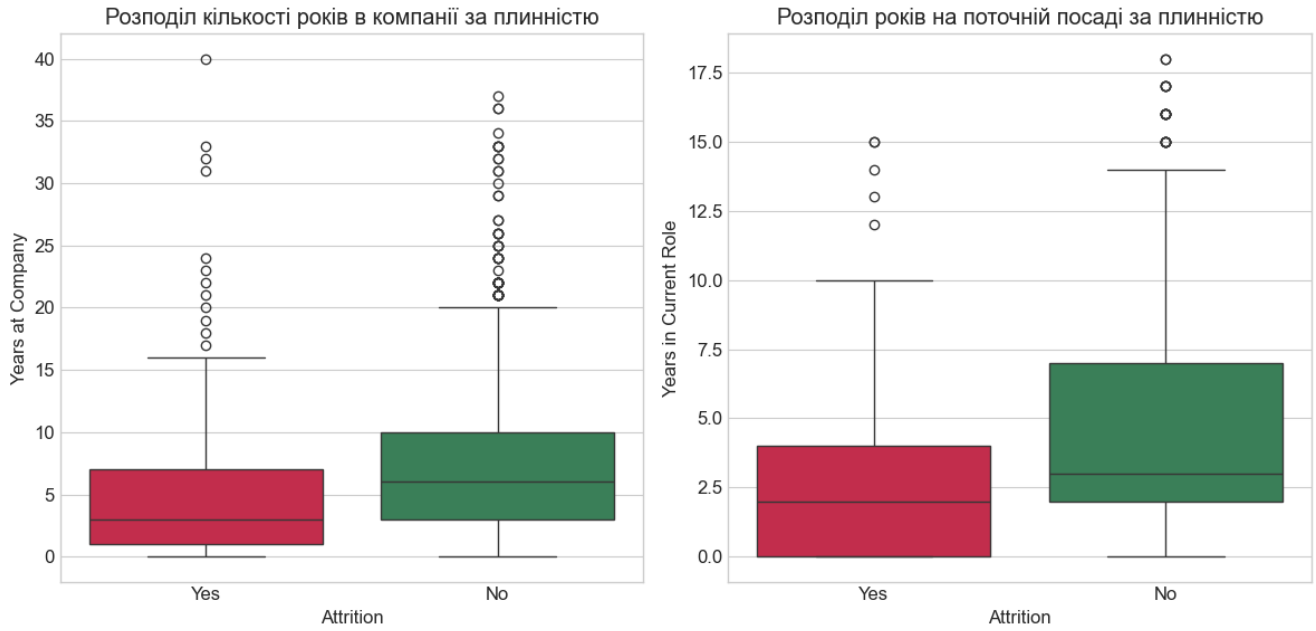

Рис. 2.15. Графік розподілу кількості років у компанії та на посаді за плинністю

**Джерело: створено автором [Додаток А, Лістинг 20]*

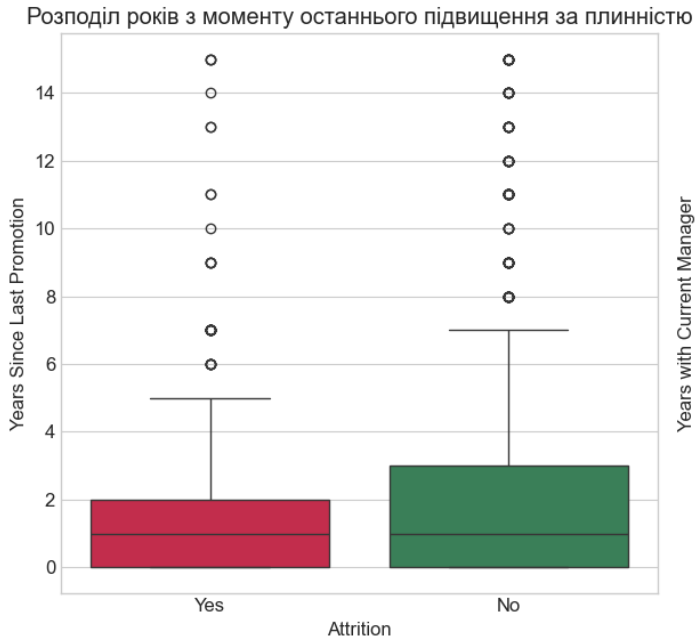

Рис. 2.16. Графік розподілу років з моменту підвищення за плинністю

**Джерело: створено автором [Додаток А, Лістинг 20]*

Цікавим є вплив фактора кількості років роботи з поточним керівником – у групі співробітників, які працювали з одним менеджером короткий час, плинність є вищою, що може свідчити про складнощі в адаптації до нового стилю керівництва. Нарешті, показник кількості попередніх місць роботи свідчить, що особи з багатим досвідом зміни компаній частіше змінюють і поточне місце праці, що можна пояснити їхньою вищою мобільністю на ринку праці (Див. Рис. 2.17).

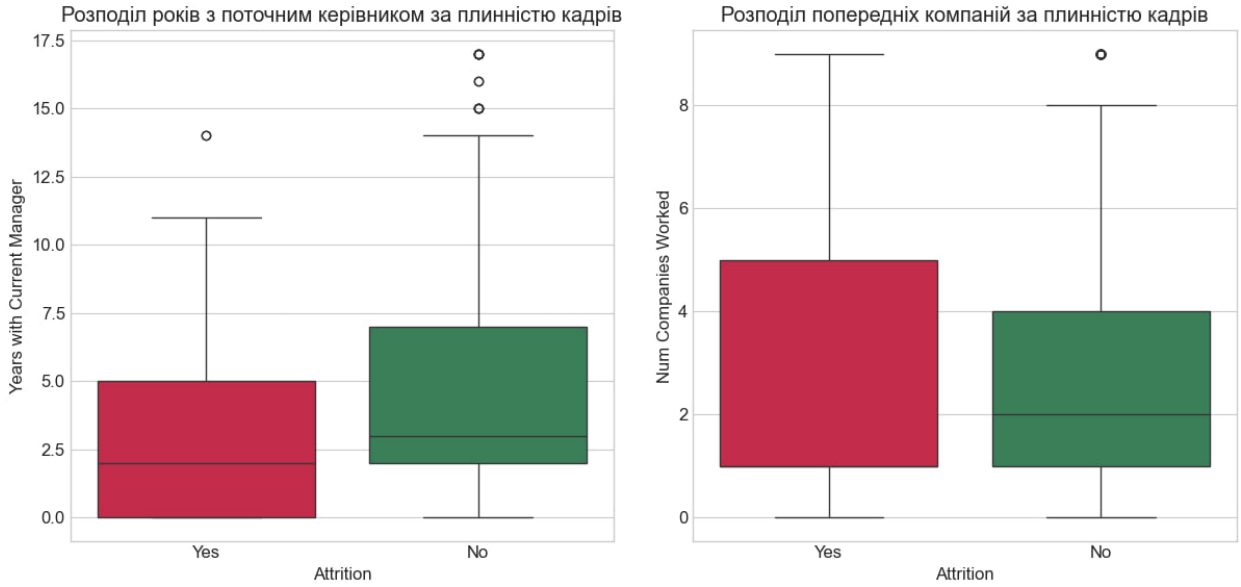

Рис. 2.17. Графік розподілу років з керівником та кількості компаній за плинністю

**Джерело: створено автором [Додаток А, Лістинг 20]*

Таким чином, досвід і стаж роботи є важливими факторами у моделюванні плинності кадрів, а отримані результати можуть бути використані для формування індивідуальних стратегій утримання співробітників залежно від їхнього кар'єрного шляху. Наприклад, зосередитися на покращені процесу онбордингу та адаптації нових працівників. Оскільки саме у них є вищий рівень плинності.

2.2.6. Аналіз фінансових показників

Фінансова компенсація є одним з факторів, що визначають рівень задоволеності співробітників та їхнє бажання залишатися в компанії. Зарплата, бонуси, премії та інші матеріальні стимули безпосередньо впливають на лояльність персоналу, хоча їхній вплив може бути як прямим, так і опосередкованим через інші фактори, зокрема баланс між роботою та особистим життям чи можливості для кар'єрного зростання.

Отримані результати свідчать, що фінансові показники мають помітний, але не завжди прямий вплив на плинність кадрів. Місячний дохід виявився одним із найбільш вагомих чинників. Серед співробітників, що покинули компанію, переважають працівники з нижчим рівнем доходу, що підтверджує гіпотезу про важливість конкурентної оплати праці для утримання персоналу (Див. Рис. 2.18). Водночас слід зазначити, що у групі з високими доходами також спостерігаються випадки звільнень, що може вказувати на вплив нефінансових факторів, таких як робочі умови чи можливості професійного розвитку.

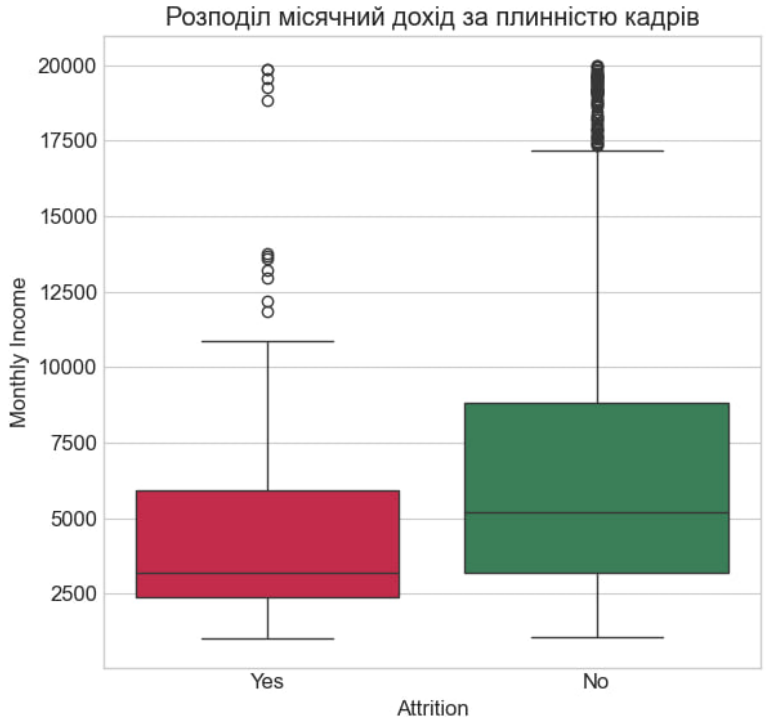

Рис. 2.18. Графік розподілу місячного доходу за плинністю

**Джерело: створено автором [Додаток А, Лістинг 21]*

Показники погодинної та денної ставки не демонструють яскраво вираженої залежності від рівня плинності, що може бути зумовлено особливостями внутрішньої системи оплати праці, де ці параметри є другорядними. Місячна ставка, як і погодинна, не виявляє значної різниці між групами (Див. Рис. 2.19).

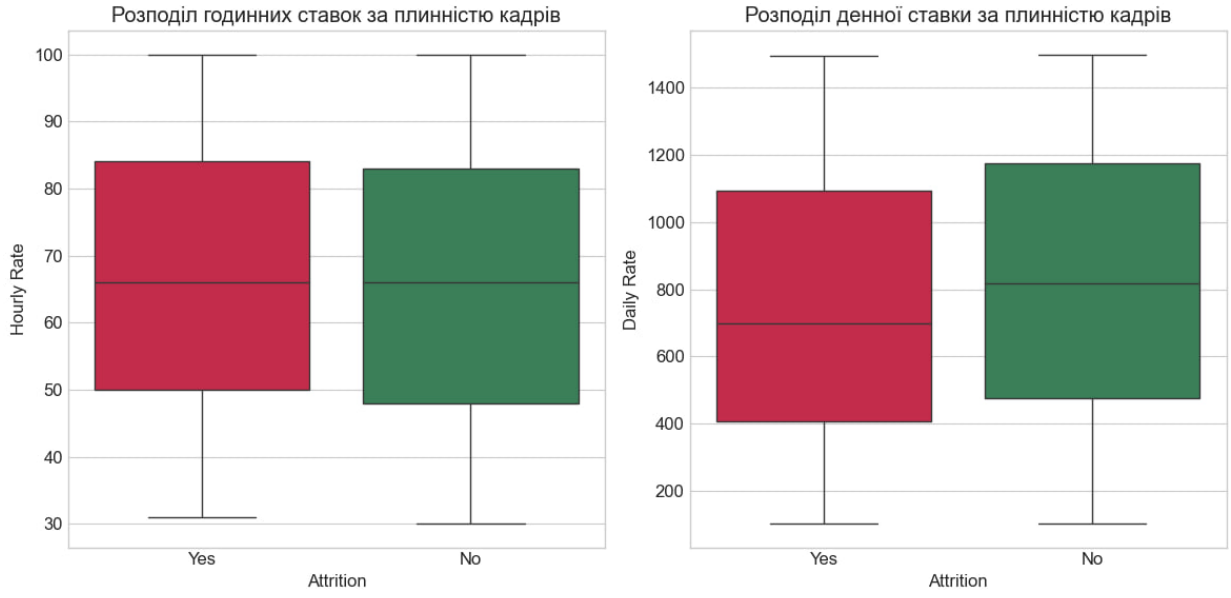

Рис. 2.19. Графіки розподілу годинної та денної ставки за плинністю

**Джерело: створено автором [Додаток А, Лістинг 21]*

Відсоток підвищення зарплати виявився показником, що має певний мотиваційний вплив. Працівники, які отримували більший відсоток підвищення, рідше залишали компанію, що підтверджує важливість регулярної індексації та фінансового заохочення (Див. Рис. 2.20).

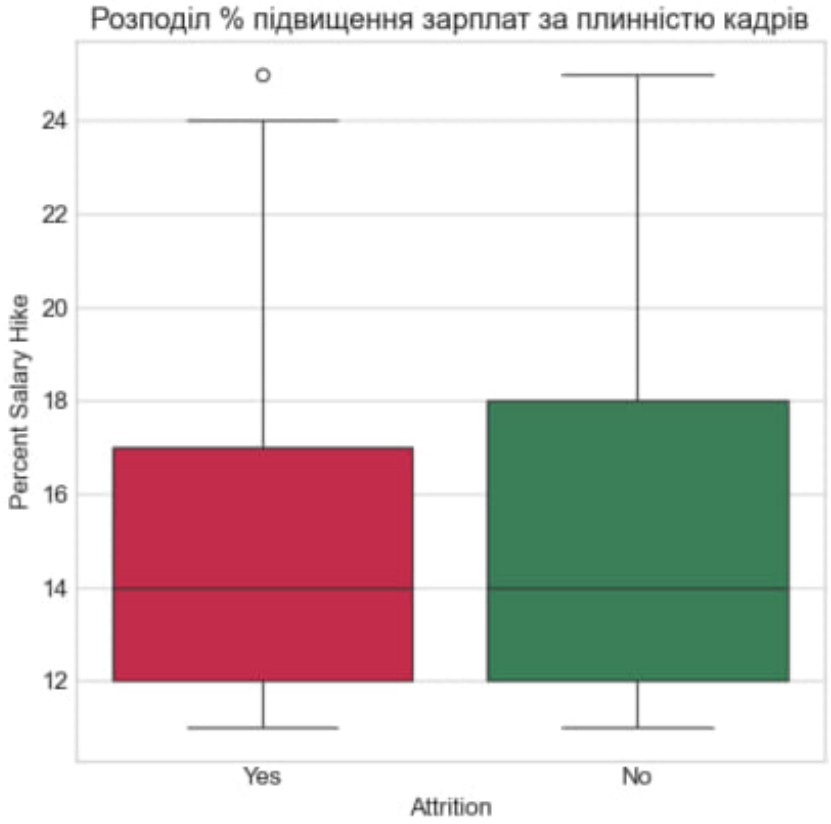

Рис. 2.20. Графік розподілу % підвищення за плинністю

**Джерело: створено автором [Додаток А, Лістинг 21]*

Щодо рівня опціонів на акції, то працівники, які мали нульовий рівень доступу до цього виду винагороди, частіше покидали компанію, ніж ті, хто мав один або кілька рівнів опціонів. Це свідчить про потенційну ефективність програм довгострокового фінансового стимулювання у зниженні плинності кадрів (Див. Рис. 2.21).

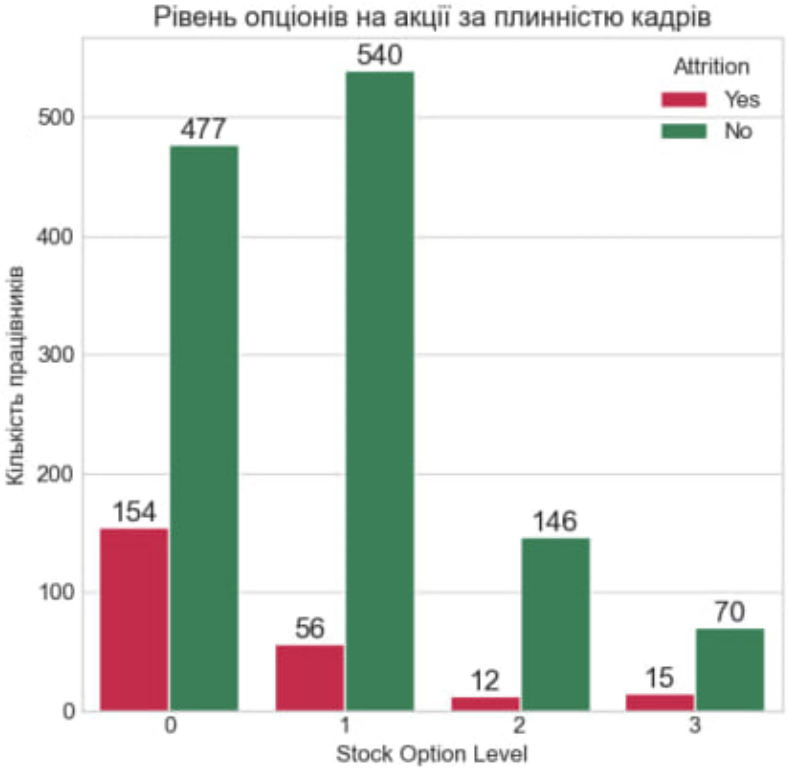

Рис. 2.21. Графік залежності відтоку від рівня опціонів

**Джерело: створено автором [Додаток А, Лістинг 21]*

2.2.7. Аналіз задоволеності робочими умовами

Робочі умови та рівень задоволеності співробітників є критичними факторами, що впливають на їхню мотивацію, продуктивність та бажання залишатися у компанії.

Аналіз показав, що понаднормові години є одним із найбільш критичних чинників, які впливають на рішення співробітників залишити компанію. Працівники, які регулярно працюють понаднормово, демонструють суттєво вищий рівень плинності кадрів, ніж ті, хто працює у межах стандартного робочого графіка. Це підтверджує важливість контролю за робочим навантаженням та запровадження гнучких графіків роботи (Див. Рис. 2.22).

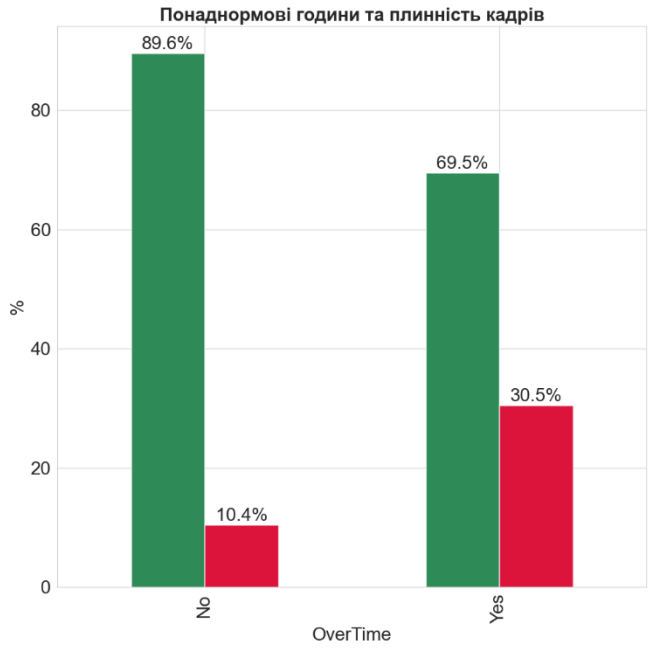

Рис. 2.22. Графік залежності відтоку від понаднормових годин

**Джерело: створено автором [Додаток А, Лістинг 22]*

Частота відряджень також виявилася помітним фактором. Найвищий рівень плинності спостерігається серед працівників, які часто подорожують у робочих цілях, що, ймовірно, пов'язано з втомою від відряджень, порушенням особистого життя та зниженням рівня задоволеності роботою. Працівники, які майже не подорожують, виявляють значно нижчу схильність до звільнень (Див. Рис. 2.23).

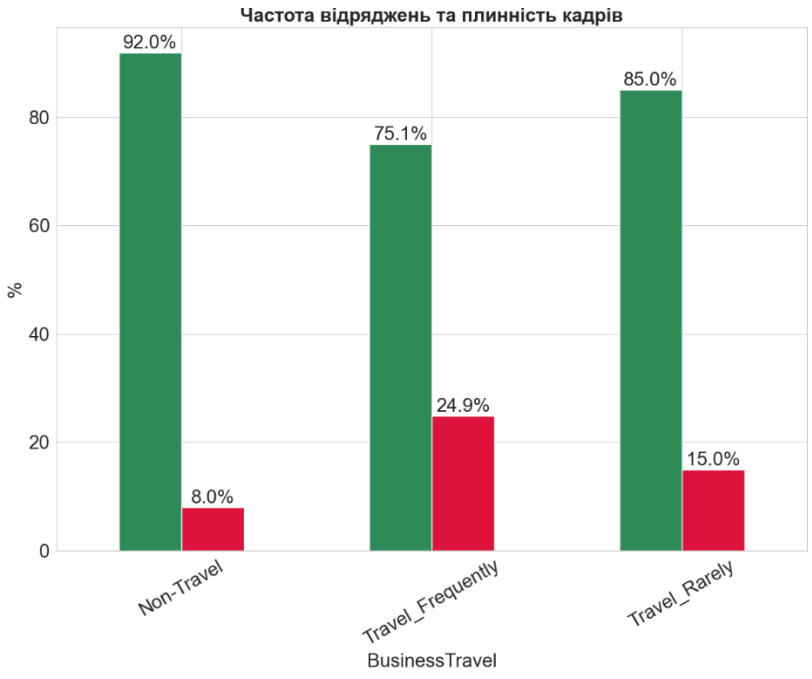

Рис. 2.23. Графік залежності відтоку від частоти відряджень та балансу роботи

**Джерело: створено автором [Додаток А, Лістинг 23]*

Показник балансу між роботою та особистим життям демонструє чітку залежність: чим вищий рівень балансу, тим нижча ймовірність плинності. Це підкреслює, що компаніям варто приділяти більше уваги політиці балансу життя та роботи, забезпечуючи співробітникам можливість гармонійно поєднувати професійну діяльність з особистими інтересами та відпочинком. Це дозволить уникнути вигорання та звільнення (Див. Рис. 2.24).

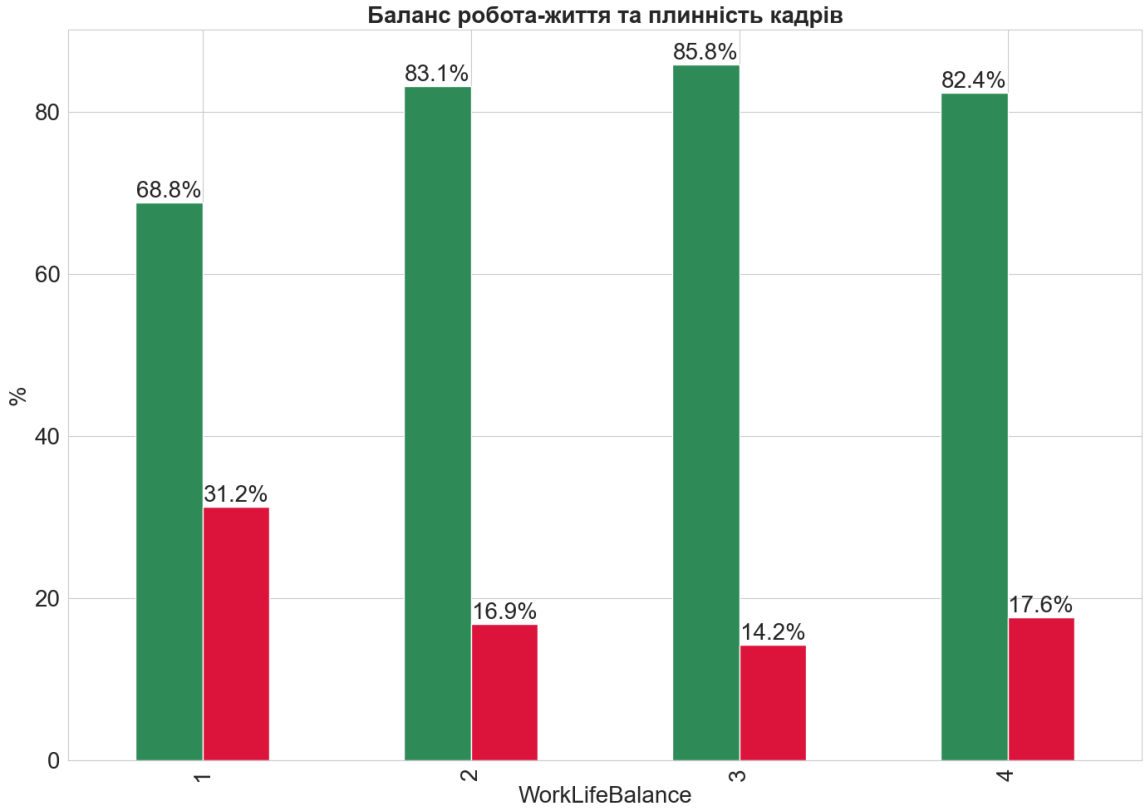

Рис. 2.24. Графік залежності відтоку від балансу роботи та життя

**Джерело: створено автором [Додаток А, Лістинг 24]*

Задоволеність робочим середовищем (Див. Рис. 2.25) і стосунками в колективі (Див. Рис. 2.26) є також вагомими аспектами. Працівники, які оцінюють ці показники на низькому рівні, значно частіше залишають компанію. Це свідчить про важливість формування сприятливого клімату в колективі, ефективного управління конфліктами та створення комфортних умов праці

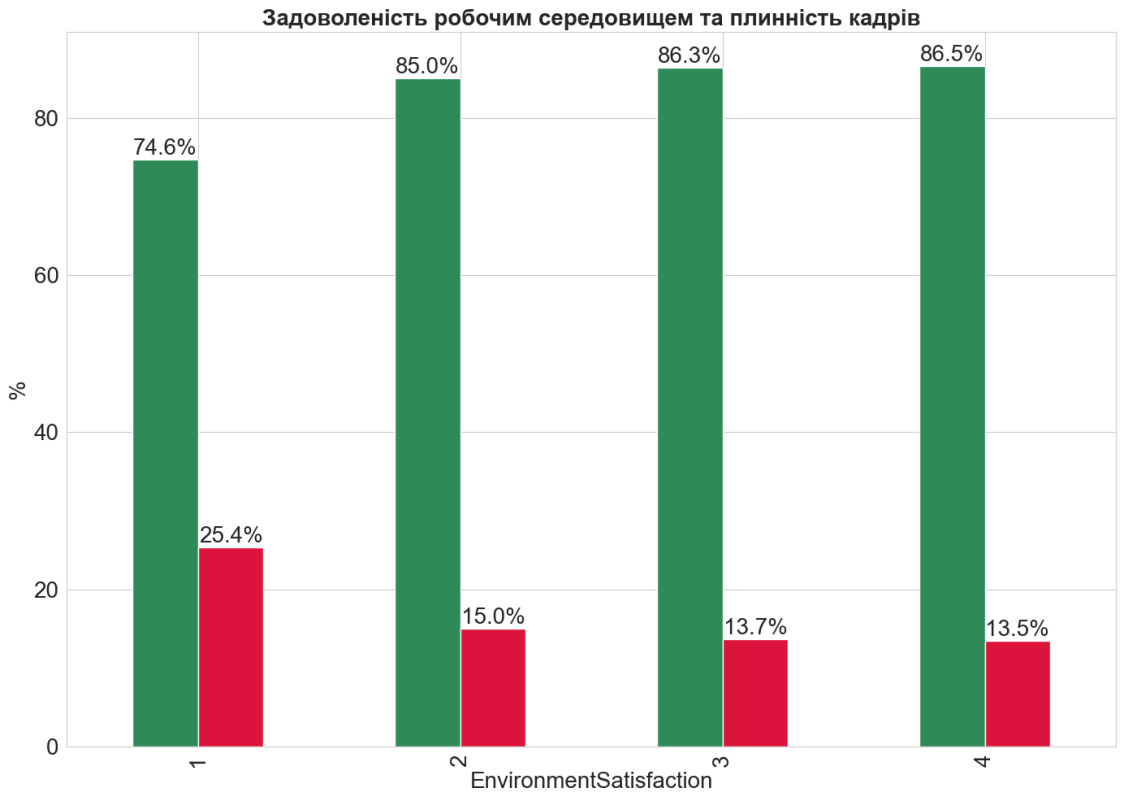

Рис. 2.25. Графік залежності відтоку від задоволеності робочим середовищем

**Джерело: створено автором [Додаток А, Лістинг 25]*

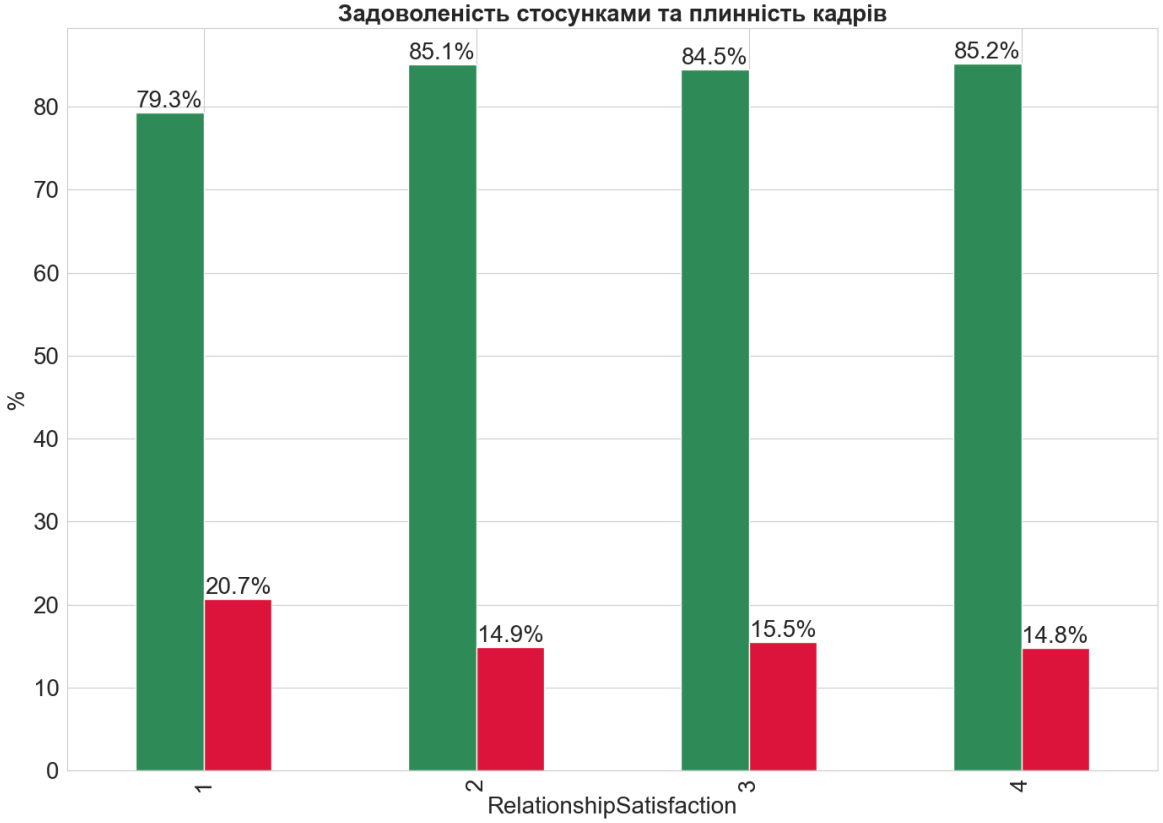

Рис. 2.26. Графік залежності відтоку від задоволеності стосунками в колективі

**Джерело: створено автором [Додаток А, Лістинг 26]*

Рівень залученості у роботу демонструє, що працівники з високим її рівнем рідше залишають компанію, оскільки вони, ймовірно, мають вищу мотивацію, відчуття значущості своєї ролі та більший рівень відповідальності. Низька залученість, навпаки, часто передує звільненню, що вказує на необхідність підтримки інтересу співробітників до їхніх завдань та проектів (Див. Рис. 2.27).

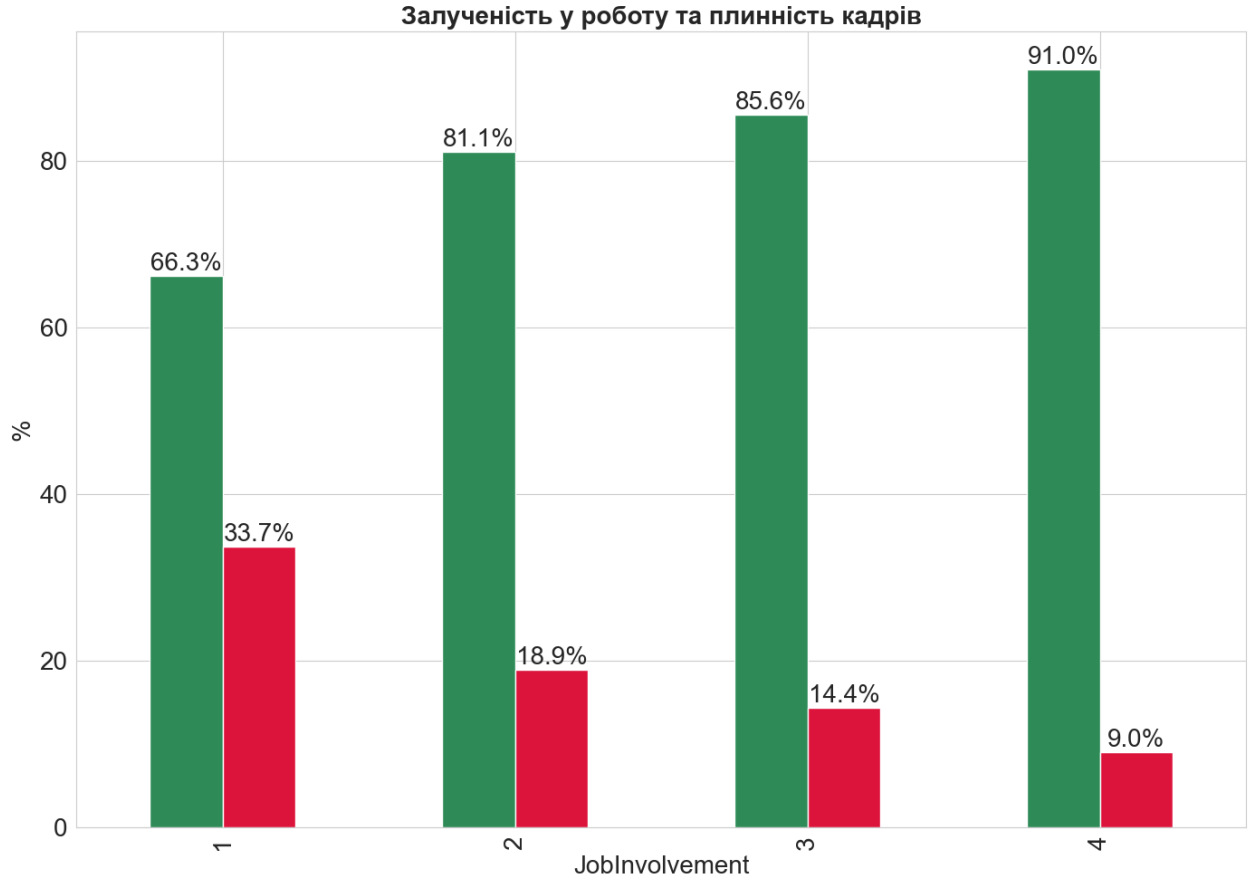

Рис. 2.27. Графік залежності відтоку від рівня залученості у роботу

**Джерело: створено автором [Додаток А, Лістинг 27]*

Таким чином, робочі умови та рівень задоволеності роботою мають комплексний вплив на плинність кадрів. Ефективна HR-стратегія повинна враховувати всі ці аспекти, забезпечуючи оптимальний баланс між навантаженням, можливостями розвитку та психологічним комфортом співробітників.

Аналіз кореляційних зв'язків між числовими змінними датасету дозволив виявити низку важливих закономірностей. Загальна кореляційна матриця показала, що значна частина змінних не має високих взаємних залежностей, що зменшує ризик мультиколінеарності у майбутньому моделюванні. Водночас спостерігаються певні очікувані взаємозв'язки: наприклад, між показниками TotalWorkingYears та

MonthlyIncome спостерігається позитивна кореляція, що логічно відображає зростання заробітної плати зі збільшенням досвіду роботи. Подібний взаємозв'язок виявлено і між змінними YearsAtCompany, YearsInCurrentRole та YearsWithCurrManager, що пояснюється природним нагромадженням досвіду у межах однієї компанії. Такі зв'язки є природними і очікуваними в HR-даних, адже відображають реальні закономірності кар'єрного розвитку працівників (Див. Рис. 2.28).

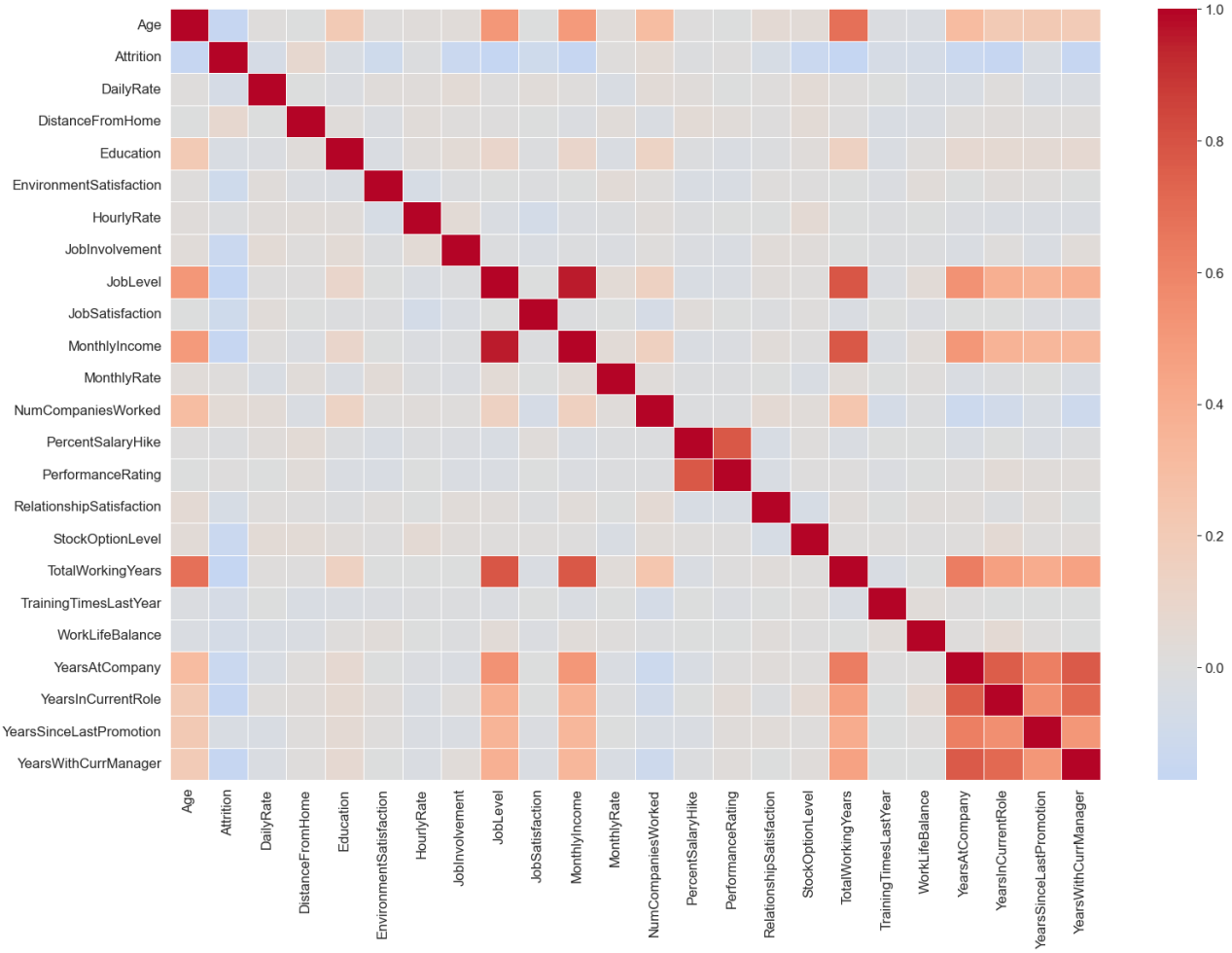

Рис. 2.28. Графік кореляції числових ознак

**Джерело: створено автором [Додаток А, Лістинг 28]*

Окремий аналіз кореляційної залежності числових характеристик із цільовою змінною Attrition дозволив визначити найбільш релевантні фактори, що можуть бути вагомими для прогнозування (Див. Рис. 2.29). Значний вплив мають змінні, пов'язані з досвідом: працівники з меншим загальним стажем або коротким періодом перебування у компанії демонструють вищу схильність до плинності. На утримання

працівників впливає рівень задоволеності, залученості та баланс життя з роботою, рівень роботи та наявність акцій компанії. Водночас змінна віддаленості роботи від дому збільшує ймовірність звільнення.

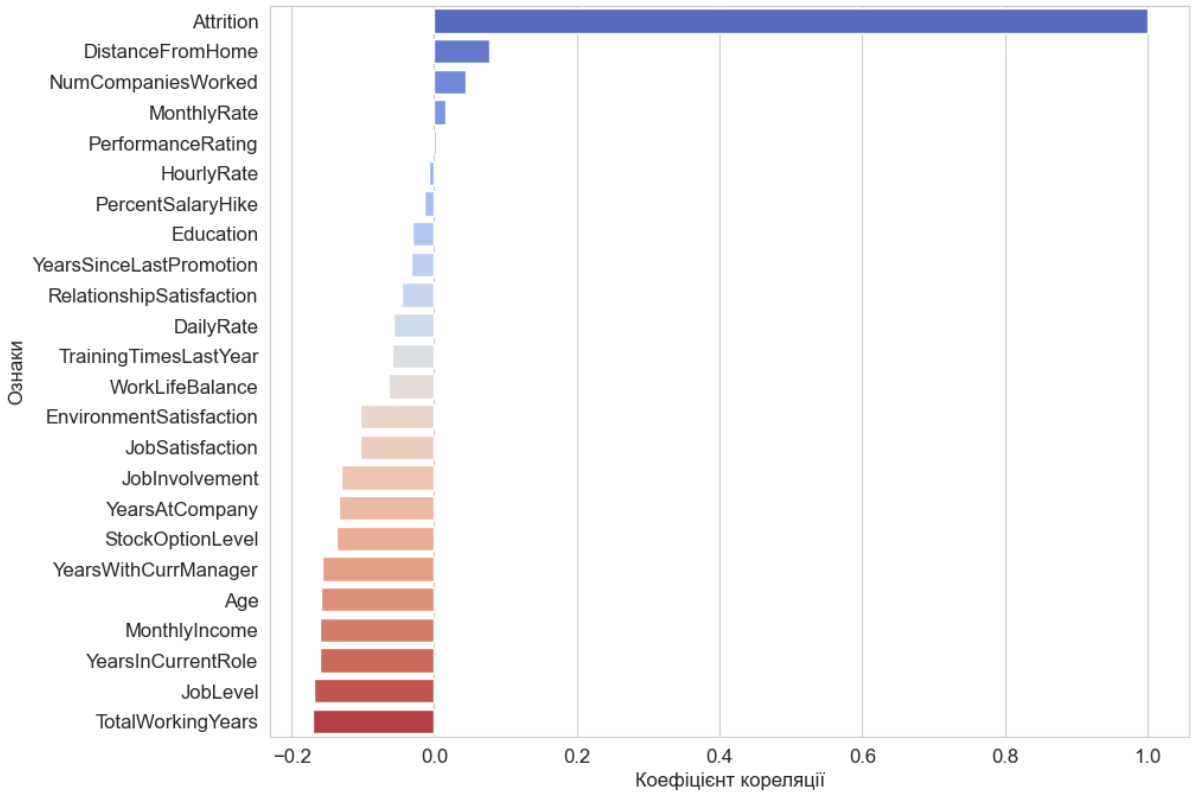

Рис. 2.29. Графік кореляції числових ознак з цільовою змінною

**Джерело: створено автором [Додаток А, Лістинг 28]*

Загалом результати кореляційного аналізу підтверджують гіпотезу про те, що фінансові фактори, показники досвіду та умови праці ключовими у формуванні ризику звільнення працівників. Це узгоджується з попередніми висновками дослідницького аналізу даних (EDA).

Отже, дослідницький аналіз даних показав, що плинність кадрів в компанії залежить від широкого спектра факторів, серед яких ключову роль відіграють вік, сімейний стан, стаж роботи, рівень задоволеності роботою, баланс між професійною та особистою сферою, а також умови праці, включаючи понаднормові години та частоту відряджень. Молодші співробітники, працівники початкових посад, особи з низькою задоволеністю роботою та слабким балансом роботи та життя, а також ті, хто часто працює понаднормово чи здійснює часті відрядження, демонструють вищу схильність до звільнення. Водночас фактори, пов’язані з фінансовими стимулами,

такими як дохід і підвищення заробітної плати, виявилися менш визначальними у порівнянні з нематеріальними аспектами роботи. Отримані результати підкреслюють необхідність комплексного підходу до управління персоналом, де нарівні з матеріальними стимулами значну увагу слід приділяти покращенню умов праці, розвитку корпоративної культури та підтримці балансу між роботою та особистим життям.

2.3. Балансування та поділ даних на тестову та тренувальну вибірки

Перш ніж перейти до безпосереднього навчання моделей прогнозування плинності кадрів, необхідно провести декілька додаткових кроків обробки даних:

- Перетворити категоріальних змінні у бінарний формат.
- Масштабувати числові значення, щоб уникнути похибок у чутливих до масштабу методів машинного навчання.
- Збалансувати похибку, щоб уникнути перенавчання моделі в сторону працівників, які залишилися в компанії. Оскільки саме вони представляють більшість в наборі даних.

Моделі машинного навчання працюють з числовими даними, тому всі категоріальні змінні у наборі даних необхідно перетворити у відповідний формат. Для цього використовується кодування ознак (One-Hot Encoding), яке створює нову змінну для кожної категорії, використовуючи бінарні значення 0 або 1.

Було застосовано кодування ознак до усіх категоріальних факторних змінних та до результуючої змінної плинності кадрів. Надалі 1 репрезентуватиме схильних до звільнення працівників, а 0 – несхильних (Див. Додаток А, Лістинг 29).

Деякі алгоритми машинного навчання чутливі до масштабів числових ознак, тому їх необхідно нормалізувати.
Для цього використовувався пакет StandardScaler, який перетворює значення так, щоб вони мали середнє значення 0 та стандартне відхилення 1 (Див. Додаток А, Лістинг 30).

Наступним критично важливим етапом після попередньої обробки даних є розділення підготовленого датасету на тренувальну та тестову вибірки. Тренувальна

вибірка призначена для навчання алгоритмів машинного навчання та оптимізації їх параметрів, тоді як тестова вибірка використовується для незалежної оцінки ефективності побудованих моделей на нових даних. Така методологія є фундаментальною для забезпечення валідності результатів дослідження та запобігання ефекту перенавчання, коли модель демонструє високу точність на навчальних даних, але втрачає здатність до узагальнення на нових спостереженнях.

В рамках даного дослідження застосовується розподіл даних у співвідношенні 70:30, що відповідає усталеній практиці в галузі машинного навчання та забезпечує оптимальний баланс між обсягом даних для навчання моделей та достатністю тестової вибірки для статистично значущої оцінки їх продуктивності. Додатково задається параметр random_state для забезпечення відтворюваності результатів різних моделей (Див. Додаток А, Лістинг 31).

Оскільки цільова змінна плинності працівників є незбалансованою (83,9% працівників залишилися, а 16,1% працівників звільнилися). Такий дисбаланс може призвести до того, що модель віддаватиме перевагу прогнозуванню більш поширеного класу, ігноруючи менш представлений. Для розв’язання цієї проблеми було застосовано метод SMOTE (Synthetic Minority Oversampling Technique), який штучно збільшує кількість прикладів меншого класу шляхом генерації нових синтетичних зразків (Див. Додаток А, Лістинг 32). Балансування виконували лише на тренувальних даних, щоб уникнути впливу штучних даних на тестову вибірку та відповідно на результати якості моделей. Після балансування ми отримали по 863 спостереження для кожного значення цільової ознаки у тренувальній вибірці (Див. Рис. 2.30).

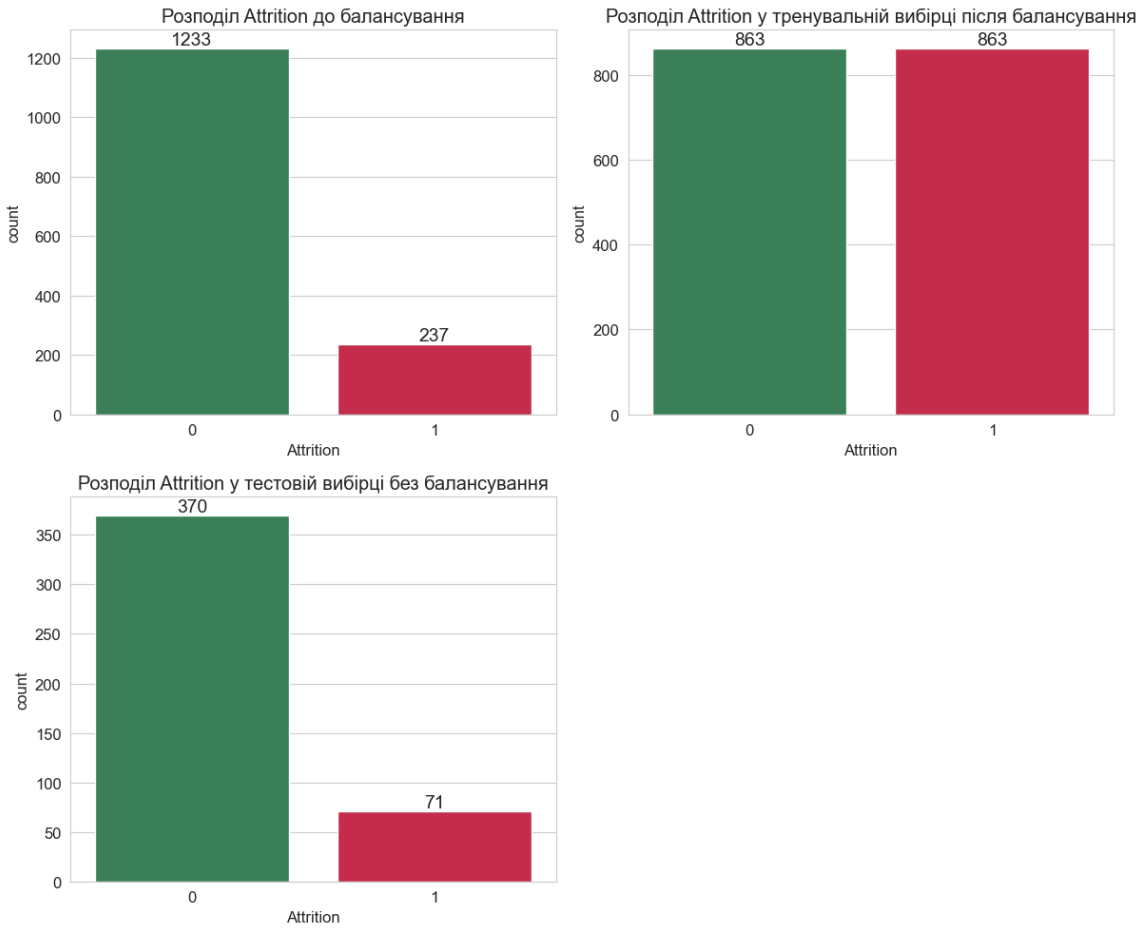

Рис. 2.30. Графіки розподілу плинності працівників до та після балансування

**Джерело: створено автором [Додаток А, Лістинг 33]*

Усі дії з підготовки даних забезпечили перетворення якісних показників у числовий формат, приведення всіх ознак до єдиного масштабу та усунення дисбалансу класів шляхом генерації синтетичних зразків для менш представленого класу. Результатом стало формування оптимізованого та збалансованого набору даних, який створює основу для подальшого етапу моделювання та розробки алгоритмів машинного навчання.

2.4. Побудова та оцінка якості моделей

Наступним етапом після дослідження та підготовки даних є створення та навчання моделей прогнозування плинності кадрів. Для цього використовували низку алгоритмів машинного навчання, які традиційно застосовуються у задачах класифікації: логістична регресія, дерево рішень, випадковий ліс, градієнтний бустинг, багатошарова нейронна мережа, алгоритм найближчих сусідів (k-NN), Naive

Bayes класифікатор та метод опорних векторів (SVM). Кожна з моделей була навчена на збалансованій тренувальній вибірці та протестована на тестовій, що дозволило більш об’єктивно оцінити їх ефективність. Для оцінки якості прогнозів використовувалися ключові метрики класифікації: Accuracy, Precision, Recall, F1-score та ROC-AUC. Такий підхід забезпечив можливість не лише обрати найкращу модель, а й зрозуміти сильні та слабкі сторони кожного алгоритму у контексті завдання передбачення звільнення працівників.

Проте варто зазначити, що у задачах прогнозування звільнень ключовим критерієм є саме Recall, оскільки він показує, яку частку працівників, що дійсно схильні до звільнення, модель здатна правильно ідентифікувати. У контексті HR-аналітики помилки другого типу False Negative описують, коли модель не розпізнає співробітника з високим ризиком звільнення. False Negative можуть призвести до значних фінансових та організаційних втрат, адже компанія втрачає кваліфікованого спеціаліста, змушена витрачати ресурси на пошук і навчання нового працівника, а також наражається на ризик зниження командної ефективності. Натомість помилки першого типу False Positive, коли модель передбачає звільнення там, де його не буде, є менш критичними, адже додаткові заходи утримання працівників можуть навіть позитивно вплинути на їхню лояльність. Тому орієнтація на високе значення Recall забезпечує більш надійну профілактику плинності кадрів і дозволяє мінімізувати ризики, пов’язані з втратою ключових співробітників.

Алгоритм дерева рішень створює ієрархічну структуру правил для класифікації об'єктів. Модель показала найвищий результат за метрикою Recall 0.52, що є особливо важливим для завдання виявлення працівників із високим ризиком звільнення. Попри нижчі значення Accuracy 0.80 та ROC-AUC 0.69, інтерпретованість моделі у вигляді дерев рішень надає можливість отримати зрозумілі пояснення щодо факторів, які впливають на рішення працівника (Див. Додаток А, Лістинг 34).

Модель випадкового лісу об'єднує множину дерев рішень і забезпечує стабільність прогнозів. Вона досягла високих показників Accuracy 0.85 та Precision 0.54, однак Recall є низьким на рівні 0.27, що знижує її придатність для завдання передбачення відтоку (Див. Додаток А, Лістинг 35).

Алгоритм градієнтного бустинг, налаштований на 200 ітерацій, продемонстрував збалансовані результати: Accuracy 0.86, Precision 0.61, але Recall залишився на низькому рівні 0.31. Це свідчить про його схильність робити більш консервативні прогнози, що зменшує чутливість до виявлення співробітників групи ризику (Див. Додаток А, Лістинг 36).

Багатошарова нейронна мережа, яка включала два приховані шари, досягла Accuracy на рівні 0.82 та Recall 0.37. Попри нелінійність та гнучкість архітектури, результати виявилися посередніми, поступаючись як логістичній регресії, так і дереву рішень у випадку задачі класифікації відтоку на нашому наборі даних (Див. Додаток А, Лістинг 37).

Алгоритм k-Nearest Neighbors показав відносно високий рівень Precision 0.57 при Accuracy 0.84, проте значення Recall залишилося низьким 0.11, що робить модель менш придатною для виявлення працівників із ризиком звільнення, оскільки модель фактично класифікує більшість працівників як тих, що не звільнилися (Див. Додаток А, Лістинг 38).

Naive Bayes класифікатор виявився несподівано ефективним за метрикою Recall 0.72, однак Accuracy 0.63 і Precision 0.27 вказують на значну кількість хибно позитивних спрацьовувань. Це обмежує можливість його практичного застосування, хоча модель може бути корисною як допоміжний інструмент для початкової ідентифікації групи ризику (Див. Додаток А, Лістинг 39).

Метод опорних векторів продемонстрував високі значення Accuracy 0.86 і Precision 0.86, але при цьому Recall залишився дуже низьким 0.17. Це означає, що модель здатна добре класифікувати більшість незвільнених працівників, але має труднощі з виявленням випадків звільнення (Див. Додаток А, Лістинг 40).

Порівняльний аналіз результатів продемонстрував, що кожен з використаних алгоритмів має свої переваги та обмеження. Логістична регресія показала найбільш збалансовані результати за ключовими метриками (Accuracy, F1-score, ROC-AUC), підтвердивши свою надійність у задачах бінарної класифікації. Випадковий ліс і градієнтний бустинг забезпечили високу точність прогнозів, проте їх чутливість (Recall) залишалася на недостатньому рівні, що знижує практичну ефективність у

завданні виявлення працівників із групи ризику. Нейронна мережа, попри складнішу архітектуру, не змогла перевершити більш прості алгоритми, що пояснюється невеликою кількістю даних та дисбалансом класів. Алгоритм k-NN показав прийнятну точність, проте його слабке значення Recall робить модель надто консервативною.

Водночас найвищий рівень Recall продемонстрував Naive Bayes класифікатор, досягнувши 0.71. Це означає, що модель здатна виявляти більшість працівників, схильних до звільнення. Проте інші метрики (Accuracy, Precision та F1-score) залишаються значно нижчими у порівнянні з іншими моделями. Така ситуація свідчить про значну кількість хибно позитивних прогнозів, що може спричинити додаткові витрати на невиправдані заходи утримання персоналу.

У цьому контексті більш оптимальним рішенням є використання дерева рішень. Попри дещо нижчі показники загальної точності, воно поєднує достатньо високий рівень Recall 0.52 із більш стабільними значеннями інших метрик, що робить його більш збалансованим варіантом. Додатковою перевагою є інтерпретованість моделі: дерево рішень дозволяє безпосередньо аналізувати правила, які впливають на прогноз, що є важливим для HR-менеджменту. Це забезпечує не лише технічну ефективність, а й практичну цінність для ухвалення управлінських рішень.

Таким чином, у процесі побудови та тестування моделей можна зробити висновок, що, хоча жодна з них не є універсальною, саме дерево рішень забезпечує найбільш прийнятний баланс між якістю прогнозів та практичною корисністю у завданні прогнозування звільнення працівників IT-компанії (Див. Рис. 2.31).

	Модель	Accuracy	Precision	Recall	F1-score	ROC-AUC
0	Random Forest	0.85	0.54	0.27	0.36	0.78
1	Decision Tree	0.80	0.42	0.52	0.46	0.69
2	Gradient Boosting	0.86	0.61	0.31	0.41	0.80
3	Нейронна мережа	0.82	0.43	0.37	0.40	0.76
4	k-NN	0.84	0.57	0.11	0.19	0.62
5	Naive Bayes	0.63	0.27	0.72	0.39	0.71
6	SVM	0.86	0.86	0.17	0.28	0.80

Рис. 2.31. Метрики усіх побудованих моделей

**Джерело: створено автором [Додаток А, Лістинг 41]*

Додатковий аналіз матриць помилок для моделей з найкращими результатами Recall дасть змогу більш детально оцінити характер їхніх помилок та здатність коректно класифікувати обидва класи цільової змінної. Особливу увагу варто звернути на Recall показник, що демонструє здатність моделі виявляти працівників, які дійсно мають намір залишити компанію (Див. Рис. 2.32).

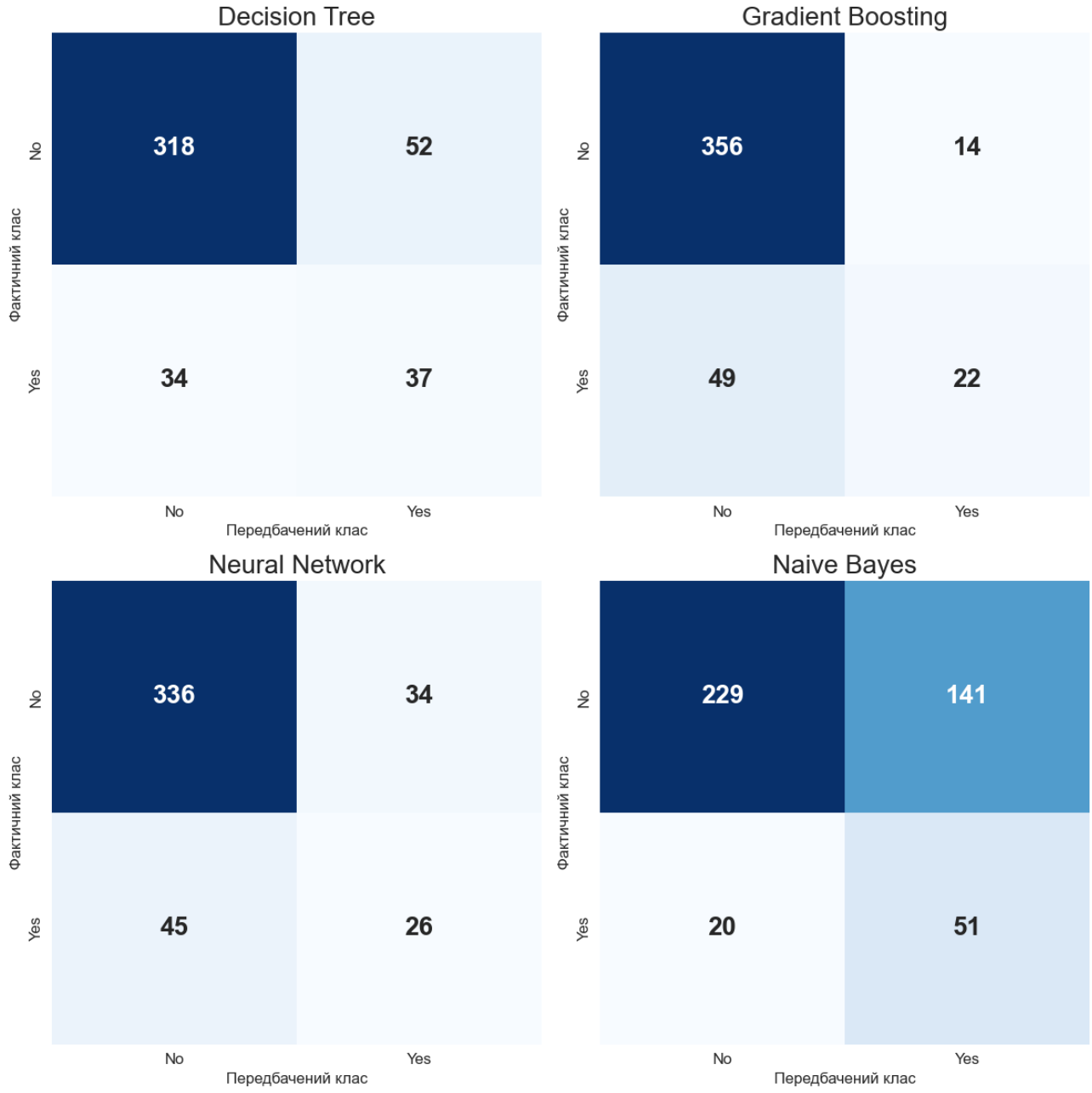

Рис. 2.32. Матриці помилок моделей з найбільшим Recall

**Джерело: створено автором [Додаток А, Лістинг 42]*

Для моделі Decision Tree матриця помилок показує, що більшість співробітників, які залишаються в компанії, класифікуються правильно, тоді як частина працівників із групи звільнення правильно виявляється лише в половині випадків. Попри це, саме ця модель продемонструвала вищий рівень Recall серед інших, що робить її придатною для задачі прогнозування плинності кадрів, адже головною метою є виявлення максимальної кількості працівників, схильних до звільнення. Можна дійти висновку, що дерево рішень забезпечило більш збалансовану роботу з менш представленим класом і може вважатися оптимальним вибором для практичного застосування у нашому випадку.

Важливою частиною побудови моделей прогнозування є не лише досягнення високих значень метрик якості, але й можливість інтерпретації результатів. Важливо розуміти, які саме фактори зумовлюють ризик звільнення працівника, оскільки це дозволяє розробляти цільові управлінські рішення. Тому у межах дослідження було проведено аналіз важливості ознак, побудовано дерево рішень для наочного представлення логіки моделі, а також проаналізовано ROC-AUC.

Для визначення важливості факторів було використано модель випадкового лісу (Random Forest), яка агрегує результати множини дерев рішень і надає кількісну оцінку значущості кожної змінної. Важливість ознак у випадковому лісі розраховується на основі того, наскільки використання певної змінної зменшує показник невизначеності під час поділу вузлів у дереві. Чим частіше ознака використовується для поділу даних і чим сильніше вона знижує невизначеність, тим вищим є її внесок у кінцевий результат. Отримані результати показали, що найбільш вагомими факторами виявилися такі показники, як наявність понаднормової роботи, наявність можливості отримати акції компанії, кількість років в компанії та рівень посади (Див. Рис. 2.33). Це підтверджує, що поєднання фінансових, організаційних та соціально-психологічних факторів формує схильність працівників до звільнення. Важливим спостереженням є те, що саме фактор понаднормової роботи посідає одне з провідних місць серед чинників, що узгоджується з результатами дослідницького аналізу даних. Для компаній це сигнал про необхідність ретельного моніторингу

робочого навантаження співробітників, оскільки воно безпосередньо впливає на звільнення.

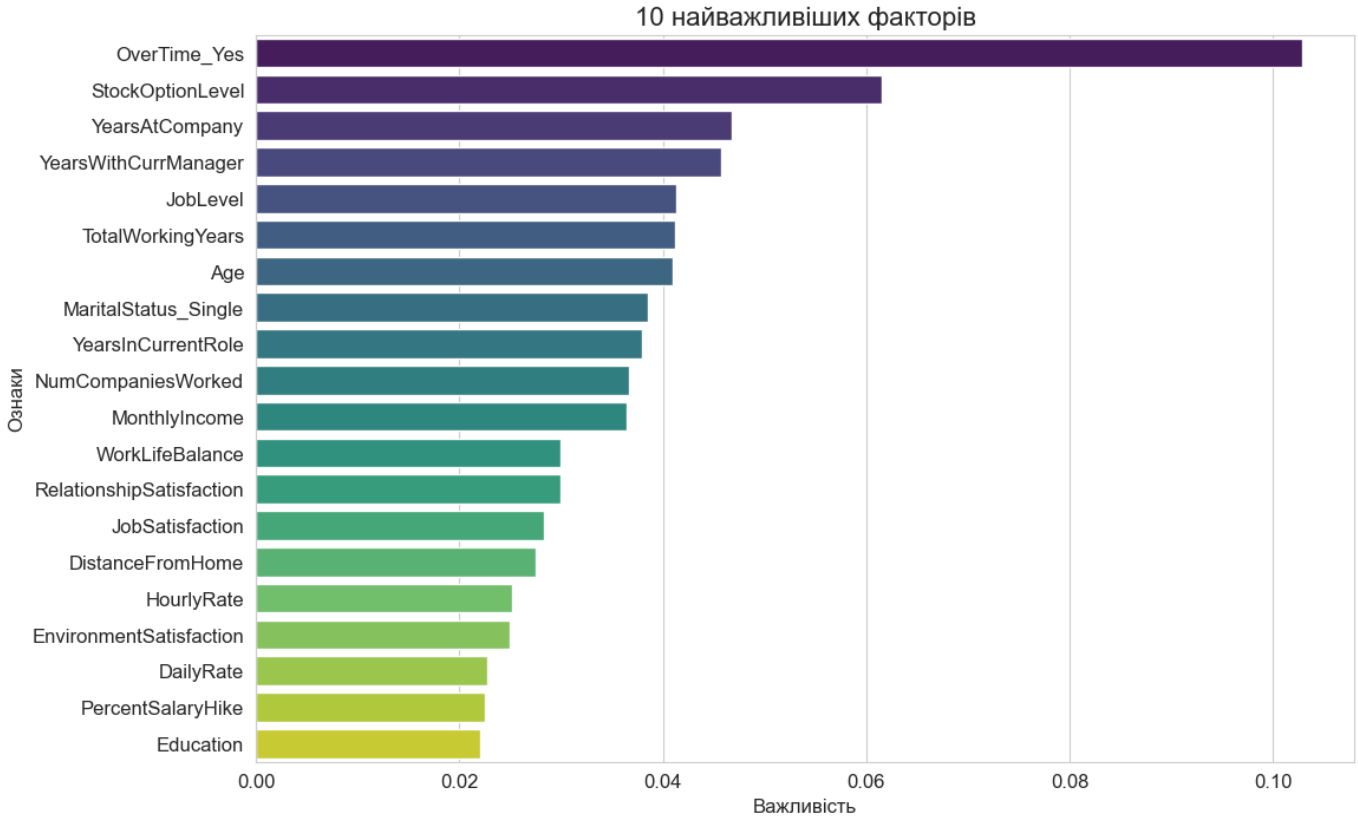

Рис. 2.33. Графік важливості ознак

**Джерело: створено автором [Додаток А, Лістинг 43]*

Щоб детальніше проілюструвати процес ухвалення рішень найкращою моделлю, було візуалізовано дерево рішень. Графічне представлення дозволяє простежити логіку, за якою алгоритм відносить конкретного працівника до групи ризику. Наприклад, модель виділяє гілки, у яких поєднання факторів «понаднормова робота = так» та «задоволеність роботою = низька» значно підвищує ймовірність звільнення. Такий підхід забезпечує зрозумілі пояснення для HR-фахівців, які можуть використовувати ці правила для формування персоналізованих заходів із підвищення лояльності персоналу. Дерево рішень у цьому випадку виконує роль не лише інструмента прогнозування, але й своєрідного діагностичного механізму для виявлення слабких місць у кадровій політиці (Див. Рис. 2.34).

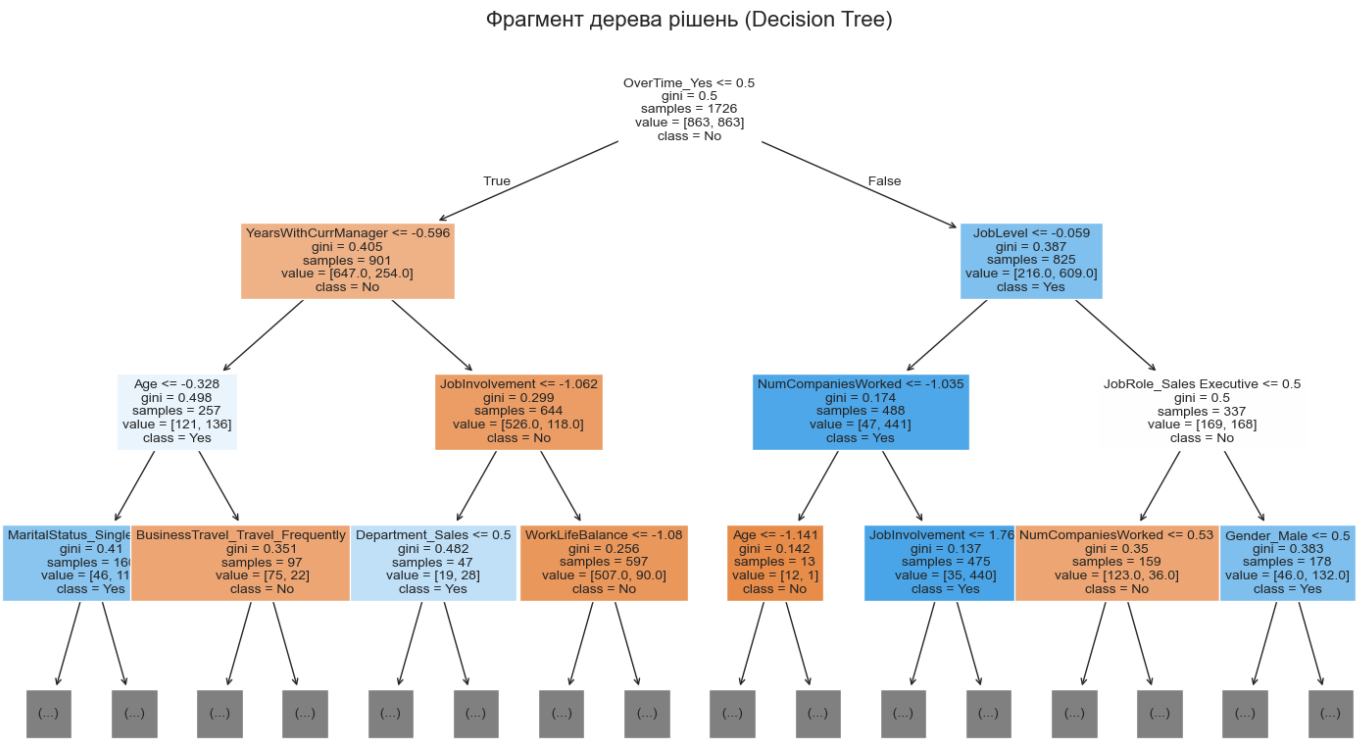

Рис. 2.34. Візуалізація дерева рішень

**Джерело: створено автором [Додаток А, Лістинг 44]*

Для кращого опису якості роботи моделі дерева рішень було побудовано ROC-криву. ROC (Receiver Operating Characteristic) крива є графічним відображенням ефективності моделей класифікації, які працюють з бінарними змінними. Вона дозволяє оцінити здатність моделі розрізняти два класи. У нашому випадку це працівники, які залишають компанію, та ті, хто продовжує працювати.

Значення площі під кривою AUC (Area Under the Curve) для моделі дерева рішень становить близько 0.69, що свідчить про достатню ефективність алгоритму у відокремленні працівників із групи ризику від тих, хто не схильний до звільнення (Див. Рис. 2.35). Чим більше значення AUC наближається до 1, тим краще модель виконує завдання класифікації. Значення, отримане у нашому випадку, є прийнятним і вказує на те, що модель здатна з відносно високою точністю визначати співробітників, які мають намір залишити компанію. Вона підтверджує придатність дерева рішень як практичного інструмента, що дає можливість компанії своєчасно виявляти працівників із підвищеною ймовірністю звільнення та реалізовувати превентивні заходи для їх утримання.

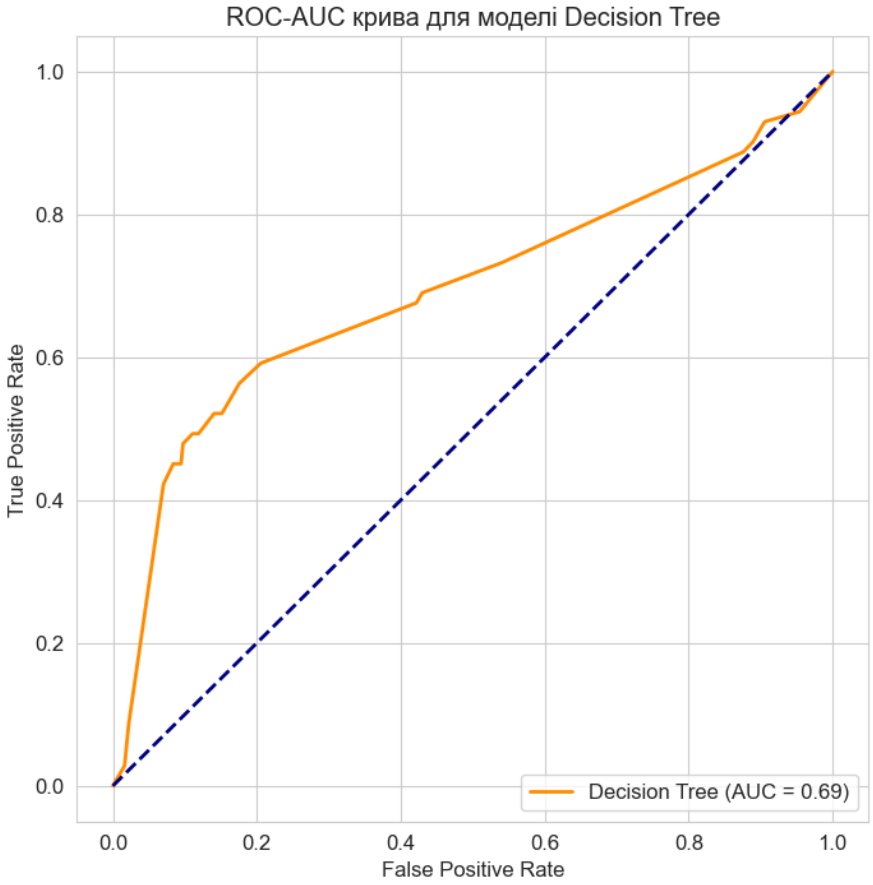

Рис. 2.35. Візуалізація ROC-AUC кривої

**Джерело: створено автором [Додаток А, Лістинг 45]*

Узагальнюючи, можна зазначити, що поєднання аналізу важливості ознак, візуалізації дерева рішень та оцінки ROC-AUC кривої забезпечує комплексне розуміння якості та інтерпретованості моделі. Це робить її придатною для практичного використання в IT-компаніях, де важливим є не лише передбачення ризику звільнення, але й отримання зрозумілих управлінських висновків.

2.5. Оцінка економічного ефекту від застосування моделі

Побудова та тестування моделей прогнозування утримання персоналу має не лише аналітичне, але й безпосереднє прикладне значення для організації. Важливим етапом є оцінка економічного ефекту, оскільки саме вона дозволяє підтвердити доцільність впровадження моделі в практику управління персоналом. Відомо, що відтік працівників супроводжується істотними витратами, пов’язаними із пошуком і підбором нових кандидатів, їх навчанням та адаптацією, а також із тимчасовим

зниженням продуктивності робочих груп. За даними HR-досліджень, вартість заміни одного фахівця IT-компанії може досягати розміру його річної заробітної плати.

На основі наявних даних було визначено річний дохід кожного працівника, який у подальшому приймався як орієнтовна величина витрат на його заміну у випадку звільнення. Загальна сума таких витрат у межах вибірки становить значну величину та відображає масштаби можливих фінансових втрат компанії за відсутності превентивних заходів (Див. Рис. 2.36).

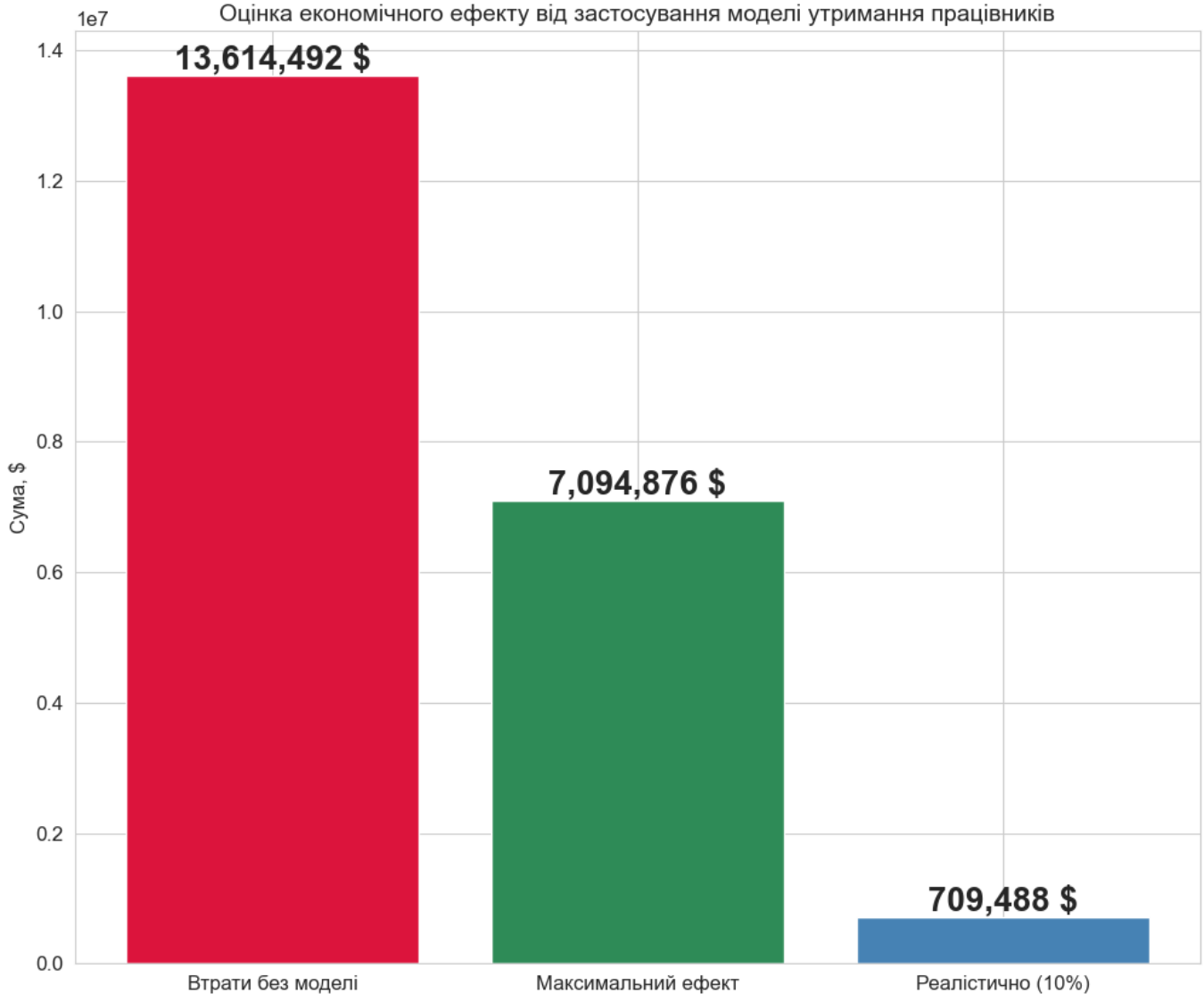

Рис. 2.36. Графік економічного ефекту моделі дерев рішень

**Джерело: створено автором [Додаток А, Лістинг 46]*

Модель дерева рішень, що показала найвищий рівень Recall (0,52), дозволяє виявити понад половину співробітників, які схильні до звільнення. Це створює

передумови для реалізації цільових заходів утримання та зменшення фактичної плинності кадрів. Розрахунок економічного ефекту було здійснено у трьох сценаріях:

- Втрати без моделі – ситуація, коли організація не використовує інструментів прогнозування і зазнає повних витрат на заміну працівників;
- Максимальний ефект – випадок, коли компанія утримує всіх працівників, виявлених моделлю;
- Реалістичний сценарій (10%) – ситуація, коли завдяки управлінським заходам вдається утримати лише 10% працівників із групи ризику, визначених моделлю.

Отримані результати демонструють суттєвий економічний ефект від впровадження моделі прогнозування плинності кадрів. У разі відсутності превентивних заходів компанія несе повні витрати на заміну співробітників, які залишають організацію. Максимальний ефект від застосування моделі дерева рішень дозволяє зменшити ці втрати більш ніж удвічі за рахунок своєчасної ідентифікації співробітників із високим ризиком звільнення. Водночас за реалістичного сценарію, коли реально вдається утримати лише 10% працівників, ідентифікованих як групу ризику для відтоку, компанія досягає відчутної економії коштів у розмірі 709 488 $. Це підтверджує практичну цінність побудованої моделі та обґрунтовує її впровадження в систему управління персоналом IT-компанії.

ВИСНОВКИ ДО РОЗДІЛУ 2

У другому розділі магістерської роботи було здійснено комплексний процес розробки та тестування моделей прогнозування плинності кадрів на основі відкритого набору даних IBM. На першому етапі провели формування та підготовку даних, виконали очищення, трансформацію змінних у зручний для моделювання формат та балансування цільової змінної за допомогою методу синтетичного збільшення вибірки. Це дало змогу уникнути зміщення моделі у бік більш чисельного класу та забезпечити об’єктивні результати.

Подальший дослідницький аналіз даних (EDA) дав змогу отримати уявлення про основні закономірності у вибірці. Зокрема, було виявлено, що найбільший вплив на звільнення працівників мають такі фактори, як рівень заробітної плати, задоволеність роботою, тривалість роботи у компанії, наявність понаднормової праці та можливості кар’єрного зростання. Після було побудовано та протестовано низку моделей машинного навчання, що традиційно застосовуються для задач класифікації.

Порівняльний аналіз показав, що кожен алгоритм має власні сильні та слабкі сторони. Naive Bayes класифікатор показав найвищий рівень Recall, але інші його показники залишалися на низькому рівні, що обмежує його практичне застосування. Оптимальний баланс між точністю та чутливістю продемонструвала модель дерева рішень, яка виявилася найпридатнішою для розв’язання поставленого завдання. Вона дозволила виявити 52% співробітників із групи ризику, зберігаючи при цьому задовільні показники загальної точності та стабільності результатів.

Таким чином, проведене дослідження емпірично підтвердило ефективність застосування методів машинного навчання для прогнозування кадрової плинності та продемонструвало практичну цінність розроблених моделей як інструментів підвищення ефективності управління персоналом.

РОЗДІЛ 3. ПРАКТИЧНЕ ВИКОРИСТАННЯ ТА ВПРОВАДЖЕННЯ МОДЕЛІ

3.1. Інтеграція моделі у процеси управління персоналом ІТ-компанії

Розроблена у попередньому розділі модель прогнозування звільнення працівників має практичний потенціал. Її інтеграція у систему HR-аналітики дозволить автоматизувати процес виявлення співробітників із високим ризиком звільнення та своєчасно застосовувати превентивні дії для їх утримання. Це забезпечує зниження витрат, пов’язаних із плинністю кадрів, підвищує рівень мотивації та залученості працівників, а також створює передумови для формування довгострокової кадрової стратегії.

3.1.1. Місце моделі у системі HR-аналітики компанії

Інтеграція моделі є логічним кроком до цифрової трансформації управління персоналом. У сучасних умовах відділи кадрів дедалі більше переходять від традиційних адміністративних функцій до ролі стратегічного партнера бізнесу, а застосування прогнозної аналітики дає можливість ухвалювати рішення, засновані на даних.

Модель прогнозування звільнень працівників виконує функцію інструменту раннього попередження, що дозволяє виявляти співробітників із високим ризиком відтоку ще до того, як вони офіційно повідомлять про намір залишити компанію. Це створює додатковий час для управлінських дій: перегляду умов праці, проведення мотиваційних заходів, коригування навантаження або формування індивідуальних планів розвитку.

У системі HR-аналітики модель займає центральне місце серед інструментів підтримки прийняття рішень. Її результати використовуються як для щоденної роботи, так і на стратегічному рівні для розробки довгострокових політик утримання персоналу. Прикладом практичного використання може бути регулярне формування списків співробітників із найвищим ризиком звільнення, що надаються HR-відділу та керівникам підрозділів.

Модель може бути застосована наступним чином:

- HR-фахівці застосовують її для оперативного аналізу ризиків та планування індивідуального підходу до працівників;
- Керівники відділів отримують змогу контролювати стабільність команди та своєчасно реагувати на потенційні кадрові проблеми;
- Вищий рівень керівників використовує агреговані показники для стратегічного управління, формування кадрової політики та визначення економічних наслідків плинності кадрів для компанії.

Отже, модель не є ізольованим інструментом, а інтегрується у багаторівневу систему, забезпечуючи різні рівні управління релевантною інформацією. Її впровадження дозволяє перейти від реактивного управління персоналом до прогнозного та превентивного підходу для підвищення ефективності кадрової стратегії.

3.1.2. Технологічна архітектура реалізації

Архітектура впровадження моделі у бізнес-процеси має ґрунтуватися на принципах модульності, масштабованості та інтегрованості з існуючими системами. Це означає, що модель не функціонує ізольовано, а вбудовується у середовище корпоративної HR-аналітики, забезпечуючи безперервний цикл збору даних, їхньої обробки, аналізу та формування прогнозів.

Основою технологічної реалізації є мова програмування Python, яка широко використовується у сфері машинного навчання завдяки великій кількості бібліотек для роботи з даними та побудови моделей.

Щоб забезпечити безперервність оновлення даних, модель має бути інтегрована з корпоративними HRM-системами (BambooHR, Workday, SAP SuccessFactors) та ERP-рішеннями (наприклад, Microsoft Dynamics 365 або Oracle ERP). Це дозволить автоматично завантажувати ключові показники про співробітників: зарплату, участь у проєктах, результати оцінювання ефективності, відвідуваність тренінгів, понаднормові години тощо [40].

Архітектура процесу роботи моделі може бути побудована за принципом ETL-пайплайну (Extract, Transform, Load):

1. Збір даних (Extract):

- автоматичне завантаження даних із HRM/ERP-систем;
- інтеграція з корпоративними системами управління проєктами (Jira, Trello, Asana) для врахування навантаження співробітників;
- підключення до систем контролю робочого часу (наприклад, Toggl або Harvest).

2. Обробка даних (Transform):

- очищення від пропусків та дублікатів;
- перетворення категоріальних змінних у числовий формат (Label Encoding, One-Hot Encoding);
- масштабування числових даних для чутливих до масштабів алгоритмів;
- балансування вибірки з використанням SMOTE для подолання дисбалансу класів.

3. Прогнозування та аналітика (Load):

- застосування навченої моделі до нових даних на постійній основі;
- формування прогнозів ризику звільнення для кожного співробітника;
- використання A/B тестів для перевірки ефективності моделі у реальному середовищі;
- створення звітів та візуалізацій у Power BI чи Tableau, щоб відстежувати прогрес та ефективність впровадження.

Наприклад, HR-відділ щомісяця отримує оновлений звіт, де виявлені працівники з високим ризиком відтоку. Наприклад, система вказує, що у департаменті розробки 10% співробітників мають підвищену ймовірність звільнення через перевищене навантаження та низьку задоволеність роботою. Керівники відділів отримують список таких співробітників і можуть своєчасно вжити заходів: зменшити кількість понаднормових годин, надати додаткові бонуси або запропонувати кар’єрний розвиток.

Таким чином, технологічна архітектура моделі забезпечує її інтеграцію у реальні бізнес-процеси компанії, дозволяючи перетворювати дані на конкретні управлінські рішення.

3.1.3. Автоматизація процесів та масштабування

Одним із ключових завдань інтеграції моделі прогнозування плинності кадрів у діяльність ІТ-компанії є забезпечення її стабільної та автоматизованої роботи. Якщо модель запускати вручну, вона перетворюється лише на інструмент аналітичних експериментів. Натомість автоматизація дозволяє зробити її частиною щоденних HR-процесів, де прогнозування відбувається без додаткових дій.

Найбільш поширеним рішенням є розгортання моделі у вигляді API-сервісу. Наприклад, за допомогою фреймворків Flask або FastAPI можна побудувати веб-сервіс, який приймає на вхід дані співробітника (вік, стаж, рівень доходу, оцінка задоволеності роботою тощо) та повертає прогноз імовірності звільнення. Такий підхід дозволяє підключити модель безпосередньо до HRM-системи компанії. У результаті HR-менеджери отримують прогнози у реальному часі. Наприклад, під час оновлення даних у профілі працівника [41].

Для забезпечення масштабованості та високої продуктивності моделі доцільним є використання хмарних рішень, таких як Microsoft Azure, Amazon Web Services (AWS), Google Cloud Platform (GCP). Ці платформи надають можливості для:

- зберігання великих обсягів даних у хмарних базах (Azure SQL Database, Amazon RDS);
- автоматизації процесів обробки даних за допомогою сервісів (Azure Data Factory, AWS Glue);
- розгортання моделей машинного навчання у вигляді готових сервісів (Azure ML, AWS SageMaker)

У такій архітектурі дані автоматично надходять у хмарне сховище, проходять необхідні трансформації, після чого модель генерує прогноз і відправляє його у HRM-систему компанії. Це дозволяє забезпечити безперервність роботи, навіть якщо обсяг даних зростає.

Практичним прикладом застосування моделі є ситуація, коли у компанії модель інтегрована як сервіс у хмарі Azure. Дані з HRM-системи автоматично надходять у Data Lake, після чого запускається щотижневий ETL-процес у Azure Data Factory. Оброблені дані подаються на вхід моделі, яка генерує список працівників із високим ризиком звільнення. Результати передаються у Power BI, де HR-відділ бачить не лише

рівень ризику, але й фактори, які найбільше вплинули на прогноз. Це дозволяє менеджерам одразу застосувати коригуючі заходи [42].

Таким чином, автоматизація роботи моделі та її масштабування забезпечують її безперервне функціонування, підвищують ефективність процесів та відкривають можливості для розширеного використання прогнозної аналітики.

3.2. Оцінка ефективності впровадження

3.2.1. Визначення ключових показників ефективності (KPI)

Для обґрунтування доцільності впровадження моделі прогнозування плинності кадрів необхідно визначити ключові показники ефективності (Key Performance Indicators), за допомогою яких можна кількісно оцінити її вплив на результати діяльності. KPI виступають об’єктивними метриками, що дозволяють виміряти успішність роботи моделі не лише з технічної, але й з економічної та організаційної точки зору.

Найважливішим показником є зниження рівня плинності кадрів. Якщо завдяки застосуванню прогнозної моделі компанія своєчасно виявляє працівників із високим ризиком звільнення та застосовує превентивні HR-заходи, фактичний відсоток звільнень зменшується. Це напряму впливає на стабільність команди, безперервність виконання проєктів та збереження фахівців [43].

Другим показником ефективності є скорочення витрат на пошук і навчання нових співробітників. Згідно з дослідженнями, заміна одного спеціаліста може коштувати компанії від 50% його річної заробітної плати [5]. Зменшення кількості звільнень за рахунок превентивних заходів дозволяє істотно знизити витрати на рекрутинг, адаптацію та навчання нових кадрів. Таким чином, модель опосередковано сприяє оптимізації фінансових потоків компанії.

Третім важливим індикатором є зростання середньої тривалості роботи співробітників у компанії. Якщо завдяки прогнозам модель допомагає утримати працівників, середній показник перебування у компанії зросте. Це означає, що в компанії знижується ймовірність зриву проєктів через кадрову нестабільність.

Важливим також є покращення якості управлінських рішень. Завдяки прогнозним даним HR-відділ та інші керівники отримують можливість ухвалювати рішення на основі об’єктивної інформації, а не лише інтуїтивних припущень. Це підвищує прозорість кадрової політики та дозволяє стратегічно планувати розвиток персоналу.

Отже, основними KPI для оцінки ефективності моделі прогнозування звільнень є: зниження плинності кадрів, скорочення витрат на підбір і навчання нових співробітників, зростання середньої тривалості роботи у компанії та підвищення якості управлінських рішень. В сукупності ці показники дають змогу об’єктивно оцінити економічну та організаційну цінність впровадження моделі в HR-процеси.

3.2.2. Економічний ефект від застосування моделі

Економічна оцінка впровадження моделі утримання персоналу має базуватися на практичних результатах роботи обраного алгоритму. У межах нашого дослідження найвищу якість прогнозування продемонструвала модель дерев рішень, яка поєднує інтерпретованість результатів із високим рівнем точності. Це забезпечує не лише технічну, а й управлінську цінність моделі, оскільки HR-фахівці та менеджери можуть чітко бачити, які саме фактори впливають на ймовірність звільнення співробітників.

Компанії зазвичай стикаються з наступними витратами на рекрутинг:

- Витрати на оплату роботи рекрутерів, використання зовнішніх платформ для пошуку кандидатів та проведення відбору;
- витрати на онбординг і навчання нових працівників, що охоплюють курси, менторську підтримку та часові витрати часу колег на адаптацію новачка;
- втрати продуктивності, пов’язані з незаповненими вакансіями та недостатнім досвідом нових співробітників у перші місяці роботи.

Застосування прогнозної моделі дозволяє значно зменшити ці витрати за рахунок зниження кількості звільнень і підвищення стабільності кадрового складу. Наприклад, навіть скорочення плинності на 5% може забезпечити економію до 15–20% HR-бюджету протягом року. Додатковим позитивним ефектом є підвищення продуктивності команд завдяки збереженню досвідчених фахівців, що у грошовому

еквіваленті може дорівнювати кільком місячним зарплатам. Оскільки найкраща модель успішно передбачає до 52% працівників з ризиком звільнення, то досягнення будь-якого з сценаріїв є цілком можливим.

Для об’єктивної оцінки ефективності доцільно використовувати сценарний аналіз.

- Оптимістичний сценарій передбачає суттєве скорочення плинності кадрів на 10-15%, що забезпечує компанії значну економію коштів та створює передумови для реінвестування у розвиток персоналу.
- Реалістичний сценарій орієнтується на поступове зниження плинності на 5–7%, що відповідає середнім очікуванням при впровадженні аналітичних інструментів. У такому випадку компанія отримує стабільний, але помірний економічний ефект.
- Песимістичний сценарій враховує можливі технічні або організаційні труднощі, які зменшують результативність моделі, обмежуючи скорочення плинності до 2–3%. У цьому випадку прямий економічний ефект буде обмеженим, однак компанія все одно отримає досвід використання прогнозних систем, що може слугувати основою для подальшого вдосконалення аналітичних процесів.

Фактично ефект від застосування моделі утримання працівників проявляється у зниженні прямих витрат на пошук і навчання персоналу, зменшенні непрямих втрат від падіння продуктивності, а також у створенні фінансових передумов для довгострокового розвитку кадрового потенціалу компанії.

3.1.4. Рекомендації щодо впровадження та покращення моделі

Впровадження моделі прогнозування утримання працівників у практику управління персоналом ІТ-компанії має відбуватися поетапно та з урахуванням як технічних, так і організаційних чинників. Головна мета полягає в тому, щоб модель стала не лише інструментом прогнозування, а й складовою системи прийняття управлінських рішень, яка підтримує стратегічні цілі компанії.

Передусім необхідно забезпечити організаційну готовність. Доцільним кроком є формування спеціалізованої робочої групи, що включатиме HR-фахівців, аналітиків

даних та спеціалістів з навчання моделей, юристів і керівників підрозділів. Така міжфункціональна команда дозволить врахувати різні аспекти впровадження. Від технічної коректності алгоритму до дотримання правових норм та етичних принципів. Важливо створити внутрішні регламенти, які визначатимуть порядок дій у разі, коли модель виявляє працівника з підвищеним ризиком звільнення. Наприклад, HR-бізнес-партнер може отримувати автоматичне сповіщення з прогнозом і поясненням факторів ризику, після чого він зобов'язаний у визначений термін провести бесіду з працівником та розробити персоналізований план утримання. Це дозволить стандартизувати процес реагування та забезпечити оперативність управлінських дій.

Окремої уваги потребує навчання HR-спеціалістів. Навіть найточніша модель може бути неефективною без належного вміння інтерпретувати її результати. Практичні тренінги мають охоплювати питання роботи з прогнозами, правильні формати комунікації з працівниками, а також основи використання інструментів пояснюваної аналітики. Це допоможе HR-персоналу краще зрозуміти, які саме фактори вплинули на прогноз, і підбирати релевантні заходи реагування.

З технологічної точки зору, доцільно впроваджувати принципи MLOps, що передбачають автоматизацію всього життєвого циклу моделі від збору та обробки даних до моніторингу її якості в реальному часі. Це дозволить вчасно виявляти зниження точності прогнозів через зміну поведінкових або організаційних чинників та своєчасно проводити перенавчання. Прикладом може бути ситуація, коли компанія запроваджує нову систему гнучких бонусів, що впливає на мотивацію персоналу. Якщо модель не враховуватиме ці зміни, її прогнози поступово втратять актуальність [44].

Особливе значення має прозорість моделі та комунікація з персоналом. Працівники мають розуміти, що система використовується не для контролю, а для створення сприятливих умов праці та підвищення рівня задоволеності. Доцільним є розроблення інформаційної кампанії, у межах якої компанія пояснюватиме мету використання моделі, принципи обробки даних та правила конфіденційності. Це дозволить знизити рівень недовіри та забезпечити позитивне сприйняття інновацій.

Серед пропозицій щодо покращення моделі варто виокремити використання ансамблевих алгоритмів, що поєднують кілька методів прогнозування для підвищення стабільності результатів. Також перспективним напрямом є застосування глибинного навчання для аналізу складніших залежностей між показниками, наприклад взаємозв'язку між результатами внутрішніх опитувань і фактичним рівнем плинності кадрів. Ще однією рекомендацією є впровадження модулів для оцінки ефективності HR-інтервенцій. Модель може фіксувати, які саме заходи застосовувалися і з яким результатом. Це створить замкнений цикл зворотного зв'язку, що дозволить підвищувати точність прогнозів і практичну користь від їх застосування.

Крім того, важливо забезпечити аудит етичних ризиків. Періодична перевірка прогнозів на наявність ознак дискримінації за статтю, віком чи іншими характеристиками дозволить своєчасно виявляти потенційні упередження. Запровадження практики «етичних звітів» щодо роботи моделі може стати додатковим інструментом підвищення довіри як з боку персоналу, так і з боку керівництва.

Ефективне впровадження та покращення моделі прогнозування утримання працівників має базуватися на інтеграції організаційних процедур, сучасних технологічних рішень та методологічного контролю. Поєднання автоматизованих прогнозів із людською експертизою, постійне вдосконалення алгоритмів та прозора комунікація із співробітниками створять оптимальні умови.

3.3. Ризики та бар'єри впровадження

3.3.1. Технічні ризики

Впровадження моделі прогнозування плинності кадрів чисто супроводжується низкою технічних ризиків, які можуть знизити її результативність або ускладнити подальшу експлуатацію. Найбільш критичним чинником є залежність від якості та повноти даних. Якщо дані HR-систем є частковими, містять пропуски, дублікати чи некоректні значення, це може призвести до викривлення прогнозів. Наприклад, відсутність інформації про кар'єрні просування працівників, понаднормову роботу,

або їхню участь у тренінгах зменшує можливість моделі правильно оцінити фактори лояльності.

Ще одним ризиком є можливість помилкових прогнозів. Жодна модель машинного навчання не забезпечує абсолютної точності. У випадку хибно позитивних прогнозів (False Positive) HR-фахівці можуть витрачати ресурси на утримання працівників, які не планували залишати компанію. Натомість хибно негативні прогнози (False Negative) несуть загрозу несподіваних звільнень ключових спеціалістів, що створює операційні проблеми.

Також важливою є потреба у регулярному оновленні моделі. Фактори, що впливають на утримання працівників, можуть змінюватися під впливом нових тенденцій. Наприклад, через поширення гібридних форматів роботи чи зростання попиту на певні спеціалізації. Без регулярного перенавчання на актуальних даних точність прогнозів поступово знижуватиметься.

Додатково слід враховувати ризики інтеграції моделі з корпоративними інформаційними системами. Технічні труднощі при поєднанні моделі з HRM або ERP-рішеннями можуть уповільнити її практичне використання. У випадку некоректної інтеграції виникає загроза затримки в отриманні прогнозів, що знижує їхню управлінську цінність [45].

Загалом технічні ризики пов’язані з якістю даних, обмеженою точністю моделі, потребою у регулярному оновленні та складністю інтеграції з корпоративними системами. Для їх мінімізації необхідне впровадження процедур контролю якості даних, системного моніторингу ефективності прогнозів і побудова гнучкої архітектури, що дозволяє своєчасно адаптувати модель до змін бізнес-середовища.

3.3.2. Організаційні бар’єри

Окрім технічних аспектів, важливим чинником успішності впровадження моделі прогнозування плинності кадрів є організаційні умови. Навіть високоточна модель, зокрема побудована на алгоритмі дерев рішень, може втратити ефективність у разі відсутності належної підтримки з боку персоналу та керівництва.

Одним із ключових бар’єрів є опір персоналу змінам. Частина співробітників може негативно сприймати використання прогнозних інструментів у сфері

управління персоналом, вважаючи це проявом надмірного контролю або втручання в особисту сферу. Це створює ризик зниження довіри до HR-практик і може призвести до погіршення корпоративного клімату, якщо процес впровадження не супроводжується належною комунікацією [46].

Іншим значним бар'єром є недостатня підготовка HR-фахівців до роботи з аналітичними моделями. Хоча дерева рішень відзначаються зрозумілою структурою та інтерпретованістю, для їх ефективного використання необхідні базові знання у сфері аналітики даних, машинного навчання та статистики. Без відповідного навчання персоналу існує ризик неправильного трактування прогнозів і, як наслідок, ухвалення неефективних управлінських рішень [46].

Також суттєвим бар'єром виступають обмежені ресурси для інтеграції та підтримки моделі. Впровадження аналітичних інструментів потребує інвестицій у технічну інфраструктуру, ліцензійне програмне забезпечення, навчання співробітників і супровід з боку спеціалістів з аналізу даних. У компаніях із обмеженим HR-бюджетом це може стати причиною відтермінування або часткової реалізації проєкту, що знижує його загальну ефективність.

Бар'єри не зводяться лише до технічної складності інтеграції моделі, а значною мірою пов'язані з людським фактором та готовністю компанії до змін. Якщо питання опору співробітників, низького рівня компетенцій HR-фахівців чи браку ресурсів залишаються невирішеними, це може звести до мінімуму потенційну користь від використання прогнозних інструментів. Тому успішність проєкту залежить не лише від обраної технології, а й від здатності організації адаптувати внутрішні процеси та культуру управління до нових підходів.

3.3.3. Етичні та правові ризики

Впровадження моделей прогнозування утримання працівників у HR-практики IT-компаній пов'язане не лише з технічними та організаційними бар'єрами, а й із низкою етичних та правових викликів. Ігнорування цих аспектів може звести нанівець позитивний ефект від використання моделі та призвести до репутаційних і фінансових втрат компанії. До них належать [47, 48]:

1. Ризик упередженості моделі. Алгоритми машинного навчання формуються на основі історичних даних, які можуть містити ознаки дискримінації. Наприклад, якщо в минулому компанія рідше просувала на керівні посади жінок чи співробітників старшого віку, то модель може успадкувати цей патерн і прогнозувати для них вищу ймовірність звільнення. У результаті HR-відділ може отримати спотворені рекомендації, що закріплюватимуть нерівність. Подібні випадки фіксувалися в міжнародній практиці. Відомим є приклад Amazon, де алгоритм відбору кандидатів віддавав перевагу чоловікам, оскільки навчався на резюме наявних співробітників компанії.

2. Питання конфіденційності та захисту персональних даних. Для побудови моделі використовуються дані про працівників: вік, стать, освіта, історія змін посад, результати опитувань, показники продуктивності. У випадку витоку чи неналежної обробки цієї інформації можливі порушення норм законодавства, зокрема Закону України «Про захист персональних даних» та регламенту про захист даних GDPR у ЄС. Наприклад, якщо компанія зберігає дані працівників у хмарних сервісах без належного шифрування, це може призвести до їх несанкціонованого доступу. Наслідками будуть штрафи, судові позови та підрив довіри.

3. Проблема прозорості та довіри. Працівники можуть сприймати модель як інструмент прихованого контролю або цифрового нагляду. Якщо HR-аналітика визначить співробітника як ризикового без пояснення причин, це здатне спричинити психологічний тиск і навіть стимулювати звільнення, замість його попередження. Наприклад, ситуація, коли менеджер повідомляє співробітнику, що алгоритм показує високий ризик вашого звільнення, але не пояснює, чому, може зруйнувати атмосферу довіри в команді.

4. Використання даних не за призначенням. Існує небезпека того, що дані, зібрані для прогнозування утримання, будуть використані для інших цілей без згоди працівників. Наприклад, для оцінки лояльності чи прийняття рішень про звільнення. Такі дії порушують принципи етичного використання даних та можуть спричинити правові наслідки.

5. Ризик надмірної автоматизації рішень. Якщо керівники повністю покладатимуться на прогноз моделі, ігноруючи індивідуальний контекст, це призведе до дегуманізації HR-процесів. Прикладом може бути ситуація, коли система визначає працівника як «нелояльного» через низький бал у внутрішньому опитуванні, хоча фактично його тимчасове невдоволення зумовлене особистими обставинами. Надмірне покладання на модель може спричинити втрату цінних кадрів.

Для мінімізації наведених ризиків доцільно впроваджувати комплекс заходів:

- проведення етичного аудиту моделей та перевірки їх на наявність дискримінаційних ознак;
- застосування методів пояснюваної аналітики, що дозволяють зрозуміти логіку прогнозів;
- формування внутрішніх політик захисту даних із чітким розподілом відповідальності;
- інформування працівників про цілі збору та використання даних, а також отримання їх згоди;
- поєднання алгоритмічних прогнозів з експертною оцінкою HR-фахівців для уникнення несправедливості.

Можна дійти висновку, що етичні та правові ризики є невід’ємною складовою впровадження прогнозних моделей у HR-аналітику. Їхнє своєчасне усвідомлення та системне управління дозволять компанії не лише дотриматися норм законодавства, а й зміцнити довіру персоналу, що є критично важливим для успішної реалізації стратегії утримання працівників.

ВИСНОВКИ ДО РОЗДІЛУ 3

У третьому розділі було представлено практичні аспекти впровадження моделі прогнозування утримання працівників в управлінську систему IT-компанії. Показано, що інтеграція такого інструменту у HR-процеси здатна істотно підвищити якість прийняття кадрових рішень, забезпечивши своєчасне виявлення ризику звільнення та створення умов для його попередження.

Аналіз ефективності моделі засвідчив, що її застосування може дати як економічний, так і соціально-організаційний ефект. Йдеться про скорочення витрат на рекрутинг і адаптацію нових працівників, зростання середньої тривалості їхньої роботи в компанії, а також зміцнення корпоративної культури й підвищення рівня задоволеності персоналу. Це, у свою чергу, формує конкурентні переваги роботодавця на ринку IT-послуг.

Разом із тим визначено низку бар'єрів і ризиків, серед яких залежність результатів від якості даних, потреба у постійному оновленні алгоритмів, можливий опір змінам з боку персоналу, а також етичні й правові виклики, пов'язані з конфіденційністю та ризиком упередженості прогнозів.

Запропоновані рекомендації акцентують на необхідності комплексного підходу. Починаючи з організаційної підготовки та навчання HR-фахівців, технологічної підтримки через сучасні практики MLOps і пояснювану аналітику, а також постійного моніторингу етичних аспектів. Така стратегія дозволить не лише підвищити точність прогнозів, а й забезпечити їхнє прийняття співробітниками та довіру до результатів.

Отже, модель прогнозування утримання працівників може стати дієвим інструментом розвитку HR-аналітики, здатним одночасно оптимізувати витрати, зміцнити лояльність персоналу та сприяти підвищенню ефективності управління людським капіталом.

ЗАГАЛЬНІ ВИСНОВКИ

У магістерській роботі було здійснено комплексне дослідження проблеми утримання працівників та запропоновано підхід до її вирішення шляхом розробки й впровадження моделі прогнозування ризику звільнення. У теоретичній частині було проаналізовано сучасні тенденції та виклики в управлінні людськими ресурсами.

У практичній частині здійснено повний цикл побудови моделі: від підготовки даних та дослідницького аналізу до навчання алгоритмів і оцінки їхньої якості. Для порівняння було протестовано кілька методів машинного навчання, серед яких дерева рішень, випадковий ліс, градієнтний бустинг, нейронні мережі та інші. Найбільш збалансовані результати продемонструвала модель дерева рішень, яка поєднує прийнятну точність прогнозів із високою інтерпретованістю.

Економічні розрахунки підтвердили практичну доцільність застосування моделі. Навіть за реалістичного сценарію, коли вдається утримати лише частину працівників із групи ризику, компанія може досягти суттєвої економії витрат на рекрутинг і навчання. Разом із тим дослідження виявило низку технічних, організаційних та етичних ризиків, серед яких залежність від якості даних, необхідність регулярного оновлення моделей, опір змінам з боку персоналу та виклики у сфері конфіденційності

Можна дійти висновку, що результати дослідження засвідчили, що інтеграція моделі прогнозування утримання працівників у систему управління персоналом є перспективним напрямом розвитку HR-аналітики. Використання такого інструменту дозволяє оптимізувати витрати, підвищити стабільність кадрового складу та зміцнити конкурентні позиції компанії. Подальші дослідження можуть бути спрямовані на вдосконалення алгоритмів шляхом використання глибинного навчання, аналізу неструктурованих даних, а також розширення сфери застосування моделі на інші HR-завдання, від прогнозування продуктивності до планування кар’єрного розвитку працівників.

СПИСОК ВИКОРИСТАН ИХ ДЖЕРЕЛ

1. Рекун, Г., & Мерінова, Н. ОСОБЛИВОСТІ УПРАВЛІННЯ РОЗВИТКОМ ПЕРСОНАЛУ В ІТ-КОМПАНІЯХ УКРАЇНИ. Молодий вчений, 3 (79), 253-257. 2020. URL: https://doi.org/10.32839/2304-5809/2020-3-79-53

2. Dino Michael Quinteros. Predictive Modelling of Employee Attrition Using Deep Learning. 2023. URL: https://doi.org/10.56578/ataiml020404

3. Chiara Morelli, Gianluca Fusai, Raffaele Zenti. Who and why will leave me? Utilizing Machine Learning-Based Models to Anticipate and Manage Employee Turnover. 2024. URL: http://dx.doi.org/10.2139/ssrn.4744130

4. Suvoj Reddy, Lydia Sri Divya. Employee Turnover Prediction - A Comparative Study of Supervised Machine Learning Models. 2022. URL: https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1679511&dswid=4512

5. Nanxi Liu, Simon Kwong Choong Mun, Alireza Mohammadi. The Impact of Colleague Departures on Employee Turnover Intentions: A Study on Chinese Enterprises. 2024. URL: https://doi.org/10.1111/ijmr.12294

6. Waterfall Methodology: A Comprehensive Guide. URL: https://www.atlassian.com/agile/project-management/waterfall-methodology

7. Agile Data Science PM. URL: https://www.datascience-pm.com/agile-data-science/

8. Kumari Jaya Kumar, V. Employee attrition forecasting: Determining the optimal algorithm for predicting employee turnover. Master's thesis, Dublin Business School. 2024. URL: https://hdl.handle.net/10788/4570

9. Nuno Rafael. Employee Turnover Predictive Model: A Model to Predict Who is More Likely to Leave a Company. 2024. URL: https://www.proquest.com/openview/e3ed294f727c1c0737639493ef7f0a0b/1

10. Mohammad Reza Shafie et al. A cluster-based human resources analytics for predicting employee turnover using optimized Artificial Neural Networks and data augmentation. 2024. URL: https://doi.org/10.1016/j.dajour.2024.100461

11. Park, J., Feng, Y. & Jeong, S.P. Developing an advanced prediction model for new employee turnover intention utilizing machine learning techniques. 2024. URL: https://doi.org/10.1038/s41598-023-50593-4

12. Gabor Szabo-Szentgroti. Modelling employee retention in small and medium-sized enterprises and large enterprises in a dynamically changing business environment. 2024. URL: https://doi.org/10.1108/IJOA-09-2023-3961

13. Schwaber, K., & Sutherland, J. (2020). The Scrum Guide: The Definitive Guide to Scrum: The Rules of the Game. URL: https://www.scrum.org/why-scrumorg

14. The CRISP-DM Process: A Comprehensive Guide. URL: https://medium.com/%40shawn.chumbar/the-crisp-dm-process-a-comprehensive-guide-4d893aecb151

15. Implementing AI for Turnover Prediction: A Step-by-Step Guide. URL: https://futuresolve.com/implementing-ai-for-turnover-prediction-a-step-by-step-guide/

16. Bersin, J., & Associates. (2022). The Future of Work in Technology. Deloitte Insights. URL: https://www.deloitte.com/us/en/insights/topics/talent/tech-leaders-reimagining-work-workforce-workplace.html

17. Bureau of Labor Statistics. (2023). Job Openings and Labor Turnover Survey. U.S. Department of Labor. URL: https://www.bls.gov/jlt/

18. Bock, L. (2018). Work Rules!: Insights from Inside Google That Will Transform How You Live and Lead. URL: https://books.google.com.ua/books/about/Work_Rules.html?id=YbPCoAEACAAJ&redir_esc=y

19. Brynjolfsson, E., & McAfee, A. (2021). The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies. URL: http://digamo.free.fr/brynmacafee2.pdf

20. Christopher M. Rosett , Austin Hagerty. Introducing HR Analytics with Machine Learning. URL: https://link.springer.com/book/10.1007/978-3-030-67626-1

21. Gelles, D. (2021). The Man Who Broke Capitalism: How Jack Welch Gutted the Heartland and Crushed the Soul of Corporate America. URL:

https://books.google.com.ua/books/about/The_Man_Who_Broke_Capitalism.html?id=8QhuEAAAQBAJ&redir_esc=y

22. Cascio, W. F., & Boudreau, J. W. (2019). Investing in People: Financial Impact of Human Resource Initiatives. URL: https://ptgmedia.pearsoncmg.com/images/9780137070923/samplepages/9780137070923.pdf

23. The Impact of AI-Driven Predictive Analytics on Employee Retention Strategies. URL: https://www.researchgate.net/publication/383791091_The_Impact_of_AI-Driven_Predictive_Analytics_on_Employee_Retention_Strategies

24. A decade of research on machine learning techniques for predicting employee turnover: A systematic literature review. URL: https://www.sciencedirect.com/science/article/abs/pii/S0957417423022960

25. Zachary Amos. 2025. Utilize machine learning to improve employee retention rates. URL: https://www.datasciencecentral.com/utilize-machine-learning-to-improve-employee-retention-rates

26. S.M. Tharani and S.V. Raj, “Predicting employee turnover intention in IT&ITeS industry using machine learning algorithms,” In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC), pp. 508-513, 2020. URL: https://ieeexplore.ieee.org/document/9243552

27. S. N. Khera and Divya, “Predictive modelling of employee turnover in Indian IT industry using machine learning techniques,” Vis. J. Bus. Perspect., vol. 23, pp. 12–21, March 2018. URL: https://www.researchgate.net/publication/331524668_Predictive_Modelling_of_Employee_Turnover_in_Indian_IT_Industry_Using_Machine_Learning_Techniques

28. K.K Mohbey, “Employee‟s Attrition Prediction Using Machine Learning Approaches,” In Machine Learning and Deep Learning in Real-Time Applications, pp. 121-128, 2020. URL: https://www.researchgate.net/publication/342401626_Employee's_Attrition_Prediction_Using_Machine_Learning_Approaches

29. D. S. Jat, P. Dhaka and A. Limbo, “Applications of statistical techniques and artificial neural networks: A review,” Journal of Statistics and Management Systems, vol. 21(4), pp. 639-645, 2018. URL: https://www.tandfonline.com/doi/abs/10.1080/09720510.2018.1475073

30. A. Singh and A. Purohit, “A survey on methods for solving data imbalance problem for classification,” Int. J. Comput. Appl., vol. 127, pp. 37–41, October 2015. URL: https://www.ijcaonline.org/archives/volume127/number15/22809-2015906677/

31. Norsuhada Mansor, Nor Samsiah Sani, Mohd Aliff. 2021. Machine Learning for Predicting Employee Attrition. URL: https://thesai.org/Downloads/Volume12No11/Paper_49-Machine_Learning_for_Predicting_Employee_Attrition.pdf

32. Filippo Guerranti, Giovanna Maria Dimitri. 2023. A Comparison of Machine Learning Approaches for Predicting Employee Attrition. URL: https://www.mdpi.com/2076-3417/13/1/267

33. Raja D V A J and Kumar R A R 2016 A Study To Reduce Employee Attrition in IT Industries International Journal of Marketing and Human Resource Management. URL: https://iaeme.com/MasterAdmin/Journal_uploads/IJMHRM/VOLUME_7_ISSUE_1/IJMHRM_07_01_001.pdf

34. Mahyar Karimi, Kamyar Seyedkazem Viliyani. 2024. Employee Turnover Analysis Using Machine Learning Algorithms. URL: https://arxiv.org/abs/2402.03905

35. Haya Alqahtani, Hana Almagrabi and Amal Alharbi. 2024. EMPLOYEE ATTRITION PREDICTION USING MACHINE LEARNING MODELS: A REVIEW PAPER. URL: https://aircconline.com/ijaia/V15N2/15224ijaia02.pdf

36. Saeed Nosratabadi, Roya Khayer Zahed. 2022. Artificial Intelligence Models and Employee Lifecycle Management: A Systematic Literature Review. URL: https://arxiv.org/abs/2209.07335

37. Довідник по Machine Learning. Gradient boosting. Itwiki. URL: https://itwiki.dev/data-science/ml-reference/ml-glossary/gradient-boosting

38. Довідник по Machine Learning. Neural Networks. Itwiki. URL: https://itwiki.dev/data-science/ml-reference/ml-glossary/neural-networks

39. Moyo, S., Doan, T. N., Yun, J. A., & Tshuma, N. (2018). Application of machine learning models in predicting length of stay among healthcare workers in underserved communities in South Africa. Human resources for health, 16(1), 1-9. https://doi.org/10.1186/s12960-018-0329-1

40. IBM HR Analytics Employee Attrition & Performance Dataset. URL: https://www.kaggle.com/datasets/pavansubhasht/ibm-hr-analytics-attrition-dataset

41. Naveen Edapurath Vijayan. 2025. Mitigating Attrition: Data-Driven Approach Using Machine Learning and Data Engineering. URL: https://arxiv.org/abs/2502.17865

42. Xiaojuan Zhu. Forecasting Employee Turnover in Large Organizations. 2016. URL: https://trace.tennessee.edu/utk_graddiss/3985/

43. Shilu Varghese, Lakshmi R. B., Geethanjali Manikrao Patil, Nikitha S. Thomas. Predictive HR Analytics: Forecasting Employee Turnover Through Machine Learning Models. URL: https://acr-journal.com/article/predictive-hr-analytics-forecasting-employee-turnover-through-machine-learning-models-1357

44. Ana Maria Căvescu, Nirvana Popescu. Predictive Analytics in Human Resources Management: Evaluating AIHR's Role in Talent Retention. URL: https://www.mdpi.com/2673-9909/5/3/99

45. Hojat Talebi, Amid Khatibi Bardsiri, Vahid Khatibi Bardsiri. Machine Learning Approaches for Predicting Employee Turnover: A Systematic Review. URL: https://onlinelibrary.wiley.com/doi/10.1002/eng2.70298

46. Oshri Bar-Gil , Tom Ron, Ofir Czerniak. AI for the people? Embedding AI ethics in HR and people analytics projects. URL: https://www.sciencedirect.com/science/article/abs/pii/S0160791X24000757

47. Laxmikant Manroop, Amina Malik, Morgan Milner. The ethical implications of big data in human resource management. URL: https://www.sciencedirect.com/science/article/abs/pii/S1053482224000020

48. Taylor Smith. The ethics of predictive HR analytics: Balancing insight with employee privacy. URL: https://www.researchgate.net/publication/390771196_THE_ETHICS_OF_PREDICTIVE_HR_ANALYTICS_BALANCING_INSIGHT_WITH_EMPLOYEE_PRIVACY

ДОДАТКИ

ДОДАТОК А

```
# Лістинг 1. Завантаження бібліотек для роботи з даними
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
# Налаштування стилю графіків
sns.set_style("whitegrid")
plt.rcParams['figure.figsize'] = [12, 8]
plt.rcParams['font.size'] = 12
colors = {'No': '#2E8B57', 'Yes': '#DC143C'}

# Лістинг 2. Завантаження даних та переегляд
data = pd.read_csv("C:\\Users\\shiki\\OneDrive\\Desktop\\Course
Paper\\IBM_HR_Attrition.csv")
# Інформація про типи даних
print(data.info())
# Розмір та перші рядки
print("Розмір датасету:", data.shape, "\n")
print(data.head())

# Лістинг 3. Описова статистика змінних
print(data.describe())

# Лістинг 4. Очищення та перевірка якості
# Перевірка на пропущені значення
print("Кількість пропущених значень:", data.isnull().sum().sum())
# Перевірка на дублікати
print("Кількість дублікатів:", data.duplicated().sum())
# Перевірка наявності константних змінних
const_cols = [col for col in data.columns if data[col].nunique() == 1]
print("Константні змінні:", const_cols)
# Видалення непотрібних змінних
data.drop(columns=const_cols + ['EmployeeNumber'], inplace=True)
print("Нова розмірність датасету:", data.shape)
# Розподіл цільової змінної
attr_counts = data['Attrition'].value_counts()
attr_perc = attr_counts / len(data) * 100
print(f"No: {attr_counts['No']} ({attr_perc['No']:.1f}%)")
print(f"Yes: {attr_counts['Yes']} ({attr_perc['Yes']:.1f}%)")

# Лістинг 5. Графіки розподілу працівників за звільненням
fig, (ax1, ax2) = plt.subplots(1, 2, figsize=(14, 8))

bars = sns.countplot(x='Attrition', data=data, hue='Attrition',
palette=[colors['Yes'], colors['No']], ax=ax1)
ax1.set_title("Кількість співробітників за фактом звільнення", fontsize=16)
ax1.set_xlabel("Attrition (Yes/No)", fontsize=16)
ax1.set_ylabel("Кількість", fontsize=16)
ax1.tick_params(labelsize=14)
for p in ax1.patches:
```

```
    height = p.get_height()
    ax1.text(p.get_x() + p.get_width()/2.,
            height + 5,
            '{:d}'.format(int(height)),
            ha="center", fontsize=18)

wedges, texts, autotexts = ax2.pie(attr_counts, labels=['Залишилися', 'Покинули'],
                                   autopct='%1.1f%%', colors=[colors['No'],
colors['Yes']])
ax2.set_title("Відсотковий розподіл працівників", fontsize=18)
plt.setp(texts, fontsize=18)
plt.setp(autotexts, fontsize=18)

plt.tight_layout()

# Лістинг 6. Розподіл віку працівників
plt.figure(figsize=(10, 10))
sns.histplot(data['Age'], bins=25, kde=True, color='#4682B4')
plt.title('Розподіл віку працівників', fontsize=18, fontweight='bold')
plt.tick_params(labelsize=18)
plt.xlabel('Вік', fontsize=18)
plt.ylabel('Кількість', fontsize=18)
plt.tight_layout()

# Лістинг 7. Розподіл віку за звільненням
plt.figure(figsize=(10, 10))
sns.histplot(data=data[data['Attrition'] == 'No'], x='Age', bins=20,
             color=colors['No'], label='Залишились', alpha=0.6)
sns.histplot(data=data[data['Attrition'] == 'Yes'], x='Age', bins=20,
             color=colors['Yes'], label='Покинули', alpha=0.6)
plt.title('Розподіл віку за звільненням', fontsize=18, fontweight='bold')
plt.legend(fontsize=18)
plt.tick_params(labelsize=18)
plt.xlabel('Вік', fontsize=18)
plt.ylabel('Кількість', fontsize=18)
plt.tight_layout()

# Лістинг 8. Відтік за віком (boxplot)
plt.figure(figsize=(10, 10))
sns.boxplot(data=data, x='Attrition', y='Age', palette=[colors['Yes'],
colors['No']])
plt.title('Відтік за віком', fontsize=18, fontweight='bold')
plt.xticks([0, 1], ['Покинули', 'Залишились'], fontsize=18)
plt.tick_params(labelsize=18)
plt.xlabel('Статус працівника', fontsize=18)
plt.ylabel('Вік', fontsize=18)
plt.tight_layout()

# Лістинг 9. Плинність за статтю
plt.figure(figsize=(10, 10))
gender_attrition = pd.crosstab(data['Gender'], data['Attrition'],
normalize='index') * 100
ax = gender_attrition.plot(kind='bar', color=[colors['No'], colors['Yes']])
plt.title('Плинність за статтю', fontsize=18, fontweight='bold')
```

```
plt.xticks([0, 1], ['Жіноча', 'Чоловіча'], rotation=0, fontsize=18)
plt.legend(['Залишились', 'Покинули'], fontsize=18)
plt.tick_params(labelsize=18)
plt.xlabel('Стать', fontsize=18)
plt.ylabel('Відсоток (%)', fontsize=18)

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontweight='bold', fontsize=12)
plt.tight_layout()

# Лістинг 10. Плинність за сімейним станом
plt.figure(figsize=(10, 10))
marital_attrition = pd.crosstab(data['MaritalStatus'], data['Attrition'],
normalize='index') * 100
ax = marital_attrition.plot(kind='bar', color=[colors['No'], colors['Yes']])
plt.title('Плинність за сімейним станом', fontsize=18, fontweight='bold')
plt.xticks([0, 1, 2], ['Розлучений', 'Одружений', 'Неодружений'], rotation=45,
fontsize=18)
plt.legend(['Залишились', 'Покинули'], fontsize=18)
plt.tick_params(labelsize=18)
plt.xlabel('Сімейний стан', fontsize=18)
plt.ylabel('Відсоток (%)', fontsize=18)

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontweight='bold', fontsize=18)
plt.tight_layout()

# Лістинг 11. Відстань від дому
plt.figure(figsize=(10, 10))
sns.boxplot(data=data, x='Attrition', y='DistanceFromHome',
           palette=[colors['Yes'], colors['No']])
plt.title('Відстань від дому', fontsize=18, fontweight='bold')
plt.xticks([0, 1], ['Покинули', 'Залишились'], fontsize=18)
plt.tick_params(labelsize=18)
plt.xlabel('Статус працівника', fontsize=18)
plt.ylabel('Відстань від дому', fontsize=18)
plt.tight_layout()

# Лістинг 12. Плинність кадрів за відділами
fig, ax = plt.subplots(figsize=(10, 10))
dept_attrition = pd.crosstab(data['Department'], data['Attrition'],
normalize='index') * 100
dept_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Плинність кадрів за відділами (%)', fontsize=18, fontweight='bold')
ax.set_xlabel('Department', fontsize=18)
ax.set_ylabel('%', fontsize=18)
ax.tick_params(axis='x', rotation=45, labelsize=18)
ax.tick_params(axis='y', labelsize=18)
ax.get_legend().remove()

for i, container in enumerate(ax.containers):
    ax.bar_label(container, fmt='%.1f%%', fontsize=18)
plt.tight_layout()
```

```
# Лістинг 13. Плинність за рівнем посади
fig, ax = plt.subplots(figsize=(20, 10))
level_attrition = pd.crosstab(data['JobLevel'], data['Attrition'],
normalize='index') * 100
level_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Плинність за рівнем посади (%)', fontsize=24, fontweight='bold')
ax.set_xlabel('JobLevel', fontsize=24)
ax.set_ylabel('%', fontsize=24)
ax.tick_params(axis='x', labelsize=24)
ax.tick_params(axis='y', labelsize=24)
ax.get_legend().remove()
# Додаємо відсотки на графік
for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=24)
plt.tight_layout()

# Лістинг 14. Плинність за рівнем задоволеності роботою
fig, ax = plt.subplots(figsize=(14, 8))
job_sat_attrition = pd.crosstab(data['JobSatisfaction'], data['Attrition'],
normalize='index') * 100
job_sat_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Плинність за рівнем задоволеності роботою (%)', fontsize=20,
fontweight='bold')
ax.set_xlabel('JobSatisfaction', fontsize=20)
ax.set_ylabel('%', fontsize=20)
ax.tick_params(axis='x', labelsize=20)
ax.tick_params(axis='y', labelsize=20)
ax.get_legend().remove()
for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=20)
plt.tight_layout()

# Лістинг 15. Плинність за рівнем залученості у роботу
fig, ax = plt.subplots(figsize=(16, 10))
involvement_attrition = pd.crosstab(data['JobInvolvement'], data['Attrition'],
normalize='index') * 100
involvement_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Плинність за рівнем залученості (%)', fontsize=24,
fontweight='bold')
ax.set_xlabel('JobInvolvement', fontsize=24)
ax.set_ylabel('%', fontsize=24)
ax.tick_params(axis='x', labelsize=24)
ax.tick_params(axis='y', labelsize=24)
ax.get_legend().remove()

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=24)
plt.tight_layout()

# Лістинг 16. Плинність за оцінкою ефективності
fig, ax = plt.subplots(figsize=(10, 10))
perf_attrition = pd.crosstab(data['PerformanceRating'], data['Attrition'],
normalize='index') * 100
```

```
perf_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Плинність за оцінкою ефективності (%)', fontsize=22,
fontweight='bold')
ax.set_xlabel('PerformanceRating', fontsize=22)
ax.set_ylabel('%', fontsize=22)
ax.tick_params(axis='x', labelsize=22)
ax.tick_params(axis='y', labelsize=22)
ax.get_legend().remove()

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=22)
plt.tight_layout()

# Лістинг 17. Плинність за балансом роботи та життя
fig, ax = plt.subplots(figsize=(14, 10))
balance_attrition = pd.crosstab(data['WorkLifeBalance'], data['Attrition'],
normalize='index') * 100
balance_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Плинність за балансом роботи та життя (%)', fontsize=22,
fontweight='bold')
ax.set_xlabel('WorkLifeBalance', fontsize=22)
ax.set_ylabel('%', fontsize=22)
ax.tick_params(axis='x', labelsize=22)
ax.tick_params(axis='y', labelsize=22)
ax.get_legend().remove()

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=22)
plt.tight_layout()
plt.show()

# Лістинг 18. Плинність за понаднормовими годинами
fig, ax = plt.subplots(figsize=(10, 10))
overtime_attrition = pd.crosstab(data['OverTime'], data['Attrition'],
normalize='index') * 100
overtime_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Плинність за понаднормовими годинами (%)', fontsize=22,
fontweight='bold')
ax.set_xlabel('OverTime', fontsize=22)
ax.set_ylabel('%', fontsize=22)
ax.tick_params(axis='x', labelsize=22)
ax.tick_params(axis='y', labelsize=22)
ax.get_legend().remove()

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=22)
plt.tight_layout()
plt.show()

# Лістинг 19. Плинність за частотою відряджень
fig, ax = plt.subplots(figsize=(12, 10))
travel_attrition = pd.crosstab(data['BusinessTravel'], data['Attrition'],
normalize='index') * 100
travel_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
```

```
ax.set_title('Плинність за частотою відряджень (%)', fontsize=22,
fontweight='bold')
ax.set_xlabel('BusinessTravel', fontsize=22)
ax.set_ylabel('%', fontsize=22)
ax.tick_params(axis='x', rotation=45, labelsize=22)
ax.tick_params(axis='y', labelsize=22)
ax.get_legend().remove()

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=22)
plt.tight_layout()

# Лістинг 20. Аналіз досвіду та стажу роботи
fig, axes = plt.subplots(2, 3, figsize=(18, 12))
fig.suptitle('Аналіз досвіду та стажу роботи', fontsize=16, fontweight='bold')
# 1. Загальний трудовий стаж
ax1 = axes[0, 0]
sns.boxplot(data=data, x='Attrition', y='TotalWorkingYears',
palette=[colors['Yes'], colors['No']], ax=ax1)
ax1.set_title('Розподіл трудового стажу за плинністю')
ax1.set_xlabel('Attrition')
ax1.set_ylabel('Total Working Years')
# 2. Роки в компанії
ax2 = axes[0, 1]
sns.boxplot(data=data, x='Attrition', y='YearsAtCompany', palette=[colors['Yes'],
colors['No']], ax=ax2)
ax2.set_title('Розподіл кількості років в компанії за плинністю')
ax2.set_xlabel('Attrition')
ax2.set_ylabel('Years at Company')
# 3. Роки на поточній посаді
ax3 = axes[0, 2]
sns.boxplot(data=data, x='Attrition', y='YearsInCurrentRole',
palette=[colors['Yes'], colors['No']], ax=ax3)
ax3.set_title('Розподіл років на поточній посаді за плинністю')
ax3.set_xlabel('Attrition')
ax3.set_ylabel('Years in Current Role')
# 4. Роки з моменту останнього підвищення
ax4 = axes[1, 0]
sns.boxplot(data=data, x='Attrition', y='YearsSinceLastPromotion',
palette=[colors['Yes'], colors['No']], ax=ax4)
ax4.set_title('Розподіл років з моменту останнього підвищення за плинністю')
ax4.set_xlabel('Attrition')
ax4.set_ylabel('Years Since Last Promotion')
# 5. Роки з поточним керівником
ax5 = axes[1, 1]
sns.boxplot(data=data, x='Attrition', y='YearsWithCurrManager',
palette=[colors['Yes'], colors['No']], ax=ax5)
ax5.set_title('Розподіл років з поточним керівником за плинністю кадрів')
ax5.set_xlabel('Attrition')
ax5.set_ylabel('Years with Current Manager')
# 6. Кількість попередніх компаній
ax6 = axes[1, 2]
sns.boxplot(data=data, x='Attrition', y='NumCompaniesWorked',
palette=[colors['Yes'], colors['No']], ax=ax6)
```

```
ax6.set_title('Розподіл попередніх компаній за плинністю кадрів')
ax6.set_xlabel('Attrition')
ax6.set_ylabel('Num Companies Worked')
plt.tight_layout()

# Лістинг 21. Аналіз фінансових показників
fig, axes = plt.subplots(2, 3, figsize=(18, 12))
fig.suptitle('Аналіз фінансових показників', fontsize=16, fontweight='bold')
# 1. Місячний дохід
ax1 = axes[0, 0]
sns.boxplot(data=data, x='Attrition', y='MonthlyIncome', palette=[colors['Yes'],
colors['No']], ax=ax1)
ax1.set_title('Розподіл місячний дохід за плинністю кадрів')
ax1.set_xlabel('Attrition')
ax1.set_ylabel('Monthly Income')
# 2. Годинна ставка
ax2 = axes[0, 1]
sns.boxplot(data=data, x='Attrition', y='HourlyRate', palette=[colors['Yes'],
colors['No']], ax=ax2)
ax2.set_title('Розподіл годинних ставок за плинністю кадрів')
ax2.set_xlabel('Attrition')
ax2.set_ylabel('Hourly Rate')
# 3. Денна ставка
ax3 = axes[0, 2]
sns.boxplot(data=data, x='Attrition', y='DailyRate', palette=[colors['Yes'],
colors['No']], ax=ax3)
ax3.set_title('Розподіл денної ставки за плинністю кадрів')
ax3.set_xlabel('Attrition')
ax3.set_ylabel('Daily Rate')
# 4. Місячна ставка
ax4 = axes[1, 0]
sns.boxplot(data=data, x='Attrition', y='MonthlyRate', palette=[colors['Yes'],
colors['No']], ax=ax4)
ax4.set_title('Розпділ місячної ставки за плинністю кадрів')
ax4.set_xlabel('Attrition')
ax4.set_ylabel('Monthly Rate')
# 5. Відсоток підвищення зарплати
ax5 = axes[1, 1]
sns.boxplot(data=data, x='Attrition', y='PercentSalaryHike',
palette=[colors['Yes'], colors['No']], ax=ax5)
ax5.set_title('Розподіл % підвищення зарплат за плинністю кадрів')
ax5.set_xlabel('Attrition')
ax5.set_ylabel('Percent Salary Hike')
# 6. Рівень опціонів на акції
ax6 = axes[1, 2]
sns.countplot(data=data, x='StockOptionLevel', hue='Attrition',
palette=[colors['Yes'], colors['No']], ax=ax6)
ax6.set_title('Рівень опціонів на акції за плинністю кадрів')
ax6.set_xlabel('Stock Option Level')
ax6.set_ylabel('Кількість працівників')
for container in ax6.containers:
    ax6.bar_label(container, fontsize=15)
plt.tight_layout()
```

```
# Лістинг 22. Понаднормові години та плинність кадрів
fig, ax = plt.subplots(1, 1, figsize=(10, 10))
overtime_attrition = pd.crosstab(data['OverTime'], data['Attrition'],
normalize='index') * 100
overtime_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Понаднормові години та плинність кадрів', fontsize=20,
fontweight='bold')
ax.set_xlabel('OverTime', fontsize=20)
ax.set_ylabel('%', fontsize=20)
ax.tick_params(axis='both', labelsize=20)

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=20)
ax.get_legend().remove()
plt.tight_layout()

# Лістинг 23. Частота відряджень та плинність кадрів
fig, ax = plt.subplots(1, 1, figsize=(12, 10))
travel_attrition = pd.crosstab(data['BusinessTravel'], data['Attrition'],
normalize='index') * 100
travel_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Частота відряджень та плинність кадрів', fontsize=20,
fontweight='bold')
ax.set_xlabel('BusinessTravel', fontsize=20)
ax.set_ylabel('%', fontsize=20)
ax.tick_params(axis='both', labelsize=20)
ax.tick_params(axis='x', rotation=30)

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=20)
ax.get_legend().remove()
plt.tight_layout()

# Лістинг 24. Баланс робота-життя та плинність кадрів
fig, ax = plt.subplots(1, 1, figsize=(14, 10))
balance_attrition = pd.crosstab(data['WorkLifeBalance'], data['Attrition'],
normalize='index') * 100
balance_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Баланс робота-життя та плинність кадрів', fontsize=20,
fontweight='bold')
ax.set_xlabel('WorkLifeBalance', fontsize=20)
ax.set_ylabel('%', fontsize=20)
ax.tick_params(axis='both', labelsize=20)

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=20)
ax.get_legend().remove()
plt.tight_layout()

# Лістинг 25. Задоволеність робочим середовищем та плинність кадрів
fig, ax = plt.subplots(1, 1, figsize=(14, 10))
env_attrition = pd.crosstab(data['EnvironmentSatisfaction'], data['Attrition'],
normalize='index') * 100
```

```
env_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Задоволеність робочим середовищем та плинність кадрів', fontsize=20,
fontweight='bold')
ax.set_xlabel('EnvironmentSatisfaction', fontsize=20)
ax.set_ylabel('%', fontsize=20)
ax.tick_params(axis='both', labelsize=20)

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=20)
ax.get_legend().remove()
plt.tight_layout()

# Лістинг 26. Задоволеність стосунками та плинність кадрів
fig, ax = plt.subplots(1, 1, figsize=(14, 10))
rel_attrition = pd.crosstab(data['RelationshipSatisfaction'], data['Attrition'],
normalize='index') * 100
rel_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Задоволеність стосунками та плинність кадрів', fontsize=20,
fontweight='bold')
ax.set_xlabel('RelationshipSatisfaction', fontsize=20)
ax.set_ylabel('%', fontsize=20)
ax.tick_params(axis='both', labelsize=20)

for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=20)
ax.get_legend().remove()
plt.tight_layout()

# Лістинг 27. Залученість у роботу та плинність кадрів
fig, ax = plt.subplots(1, 1, figsize=(14, 10))
involvement_attrition = pd.crosstab(data['JobInvolvement'], data['Attrition'],
normalize='index') * 100
involvement_attrition.plot(kind='bar', ax=ax, color=[colors['No'], colors['Yes']])
ax.set_title('Залученість у роботу та плинність кадрів', fontsize=20,
fontweight='bold')
ax.set_xlabel('JobInvolvement', fontsize=20)
ax.set_ylabel('%', fontsize=20)
ax.tick_params(axis='both', labelsize=20)
for container in ax.containers:
    ax.bar_label(container, fmt='%.1f%%', fontsize=20)
ax.get_legend().remove()
plt.tight_layout()

# Лістинг 28. Графік кореляції ознак
data_corr = data.copy()
data_corr['Attrition'] = data_corr['Attrition'].map({'No': 0, 'Yes': 1})
# Вибираємо тільки числові колонки
numeric_data = data_corr.select_dtypes(include=['int64', 'float64'])
# 1. Загальна матриця кореляцій
plt.figure(figsize=(18,12))
corr_matrix = numeric_data.corr()

sns.heatmap(
    corr_matrix,
```

```
    cmap="coolwarm",
    center=0,
    annot=False,
    linewidths=0.5
)
# 2. Кореляція з цільовою змінною Attrition
plt.figure(figsize=(10,8))
attrition_corr = corr_matrix['Attrition'].sort_values(ascending=False)

sns.barplot(x=attrition_corr.values, y=attrition_corr.index, palette="coolwarm")
plt.xlabel("Коефіцієнт кореляції")
plt.ylabel("Ознаки")

# Лістинг 29. Кодування ознак
from sklearn.preprocessing import LabelEncoder
df_model = data.copy()

# Кодування цільової змінної
le = LabelEncoder()
df_model['Attrition'] = le.fit_transform(df_model['Attrition'])  # No=0, Yes=1

# One-Hot Encoding для категоріальних змінних
categorical_cols = df_model.select_dtypes(include=['object']).columns
df_model = pd.get_dummies(df_model, columns=categorical_cols, drop_first=True)

print(f"Розмір датасету після кодування: {df_model.shape}")
df_model.head()

# Лістинг 30. Масштабування ознак

from sklearn.preprocessing import StandardScaler

# Визначаємо числові змінні
numeric_cols = df_model.select_dtypes(include=['int64', 'float64']).columns
numeric_cols = numeric_cols.drop('Attrition')  # виключаємо цільову змінну

# Масштабування
scaler = StandardScaler()
df_model[numeric_cols] = scaler.fit_transform(df_model[numeric_cols])

# Лістинг 31. Поділ набору даних на тестову та тренувальні вибірки
from sklearn.model_selection import train_test_split
y = df_model['Attrition']
X = df_model.drop('Attrition', axis=1)
# Розбиття на тренувальну та тестову вибірки (70% / 30%)
X_train, X_test, y_train, y_test = train_test_split(
    X, y,
    test_size=0.3,
    random_state=42,
    stratify=y
)
print(f"Тренувальна вибірка: {X_train.shape}, Тестова вибірка: {X_test.shape}")
print(f"Розподіл класів у тренувальній:\n{y_train.value_counts()}")
print(f"Розподіл класів у тестовій:\n{y_test.value_counts()}")
```

```
# Лістинг 32. Балансування цільової змінної
from imblearn.over_sampling import SMOTE
smote = SMOTE(random_state=42)
X_train_bal, y_train_bal = smote.fit_resample(X_train, y_train)
print(f"\nПісля балансування тренувальна вибірка: {X_train_bal.shape}")
print(f"Розподіл класів після балансування:\n{y_train_bal.value_counts()}")

# Лістинг 33. Графік розподілу цільової змінни до та після балансування
df_full = pd.DataFrame({'Attrition': y})              # до балансування
df_train = pd.DataFrame({'Attrition': y_train_bal})       # тренувальна після
балансування
df_test = pd.DataFrame({'Attrition': y_test})         # тестова (небалансована)
fig, axes = plt.subplots(2, 2, figsize=(12, 10))
axes = axes.flatten()
# 1. до балансування
sns.countplot(
    x='Attrition', data=df_full, hue='Attrition',
    palette=[colors['No'], colors['Yes']], ax=axes[0], legend=False
)
axes[0].set_title("Розподіл Attrition до балансування")
for container in axes[0].containers:
    axes[0].bar_label(container, labels=[f'{int(v)}' for v in
container.datavalues], fontsize=14)
# 2. Тренувальна після балансування
sns.countplot(
    x='Attrition', data=df_train, hue='Attrition',
    palette=[colors['No'], colors['Yes']], ax=axes[1], legend=False
)
axes[1].set_title("Розподіл Attrition у тренувальній вибірці після балансування")
for container in axes[1].containers:
    axes[1].bar_label(container, labels=[f'{int(v)}' for v in
container.datavalues], fontsize=14)
# 3. Тестова вибірка без балансування
sns.countplot(
    x='Attrition', data=df_test, hue='Attrition',
    palette=[colors['No'], colors['Yes']], ax=axes[2], legend=False
)
axes[2].set_title("Розподіл Attrition у тестовій вибірці без балансування")
for container in axes[2].containers:
    axes[2].bar_label(container, labels=[f'{int(v)}' for v in
container.datavalues], fontsize=14)
fig.delaxes(axes[3])
plt.tight_layout()

# Побудова моделей машинного навчання.
# Лістинг 34. Побудова та оцінка моделі дерев рішень
from sklearn.tree import DecisionTreeClassifier
dt_model = DecisionTreeClassifier(max_depth=5, random_state=42)
dt_model.fit(X_train_bal, y_train_bal)
from sklearn.metrics import accuracy_score, precision_score, recall_score,
f1_score, roc_auc_score
# Прогнозування
y_pred_dt = dt_model.predict(X_test)
```

```
y_proba_dt = dt_model.predict_proba(X_test)[:, 1]
accuracy_dt = accuracy_score(y_test, y_pred_dt)
precision_dt = precision_score(y_test, y_pred_dt)
recall_dt = recall_score(y_test, y_pred_dt)
f1_dt = f1_score(y_test, y_pred_dt)
roc_auc_dt = roc_auc_score(y_test, y_proba_dt)
print("Метрики Decision Tree:")
print(f"Accuracy: {accuracy_dt:.2f}")
print(f"Precision: {precision_dt:.2f}")
print(f"Recall: {recall_dt:.2f}")
print(f"F1-score: {f1_dt:.2f}")
print(f"ROC-AUC: {roc_auc_dt:.2f}")

# Лістинг 35. Побудова та оцінка моделі випадкових лісів
from sklearn.ensemble import RandomForestClassifier
rf_model = RandomForestClassifier(n_estimators=200, max_depth=10, random_state=42)
rf_model.fit(X_train_bal, y_train_bal)
# Прогнозування
y_pred_rf = rf_model.predict(X_test)
y_proba_rf = rf_model.predict_proba(X_test)[:, 1]
accuracy_rf = accuracy_score(y_test, y_pred_rf)
precision_rf = precision_score(y_test, y_pred_rf)
recall_rf = recall_score(y_test, y_pred_rf)
f1_rf = f1_score(y_test, y_pred_rf)
roc_auc_rf = roc_auc_score(y_test, y_proba_rf)
print("Метрики Random Forest:")
print(f"Accuracy: {accuracy_rf:.2f}")
print(f"Precision: {precision_rf:.2f}")
print(f"Recall: {recall_rf:.2f}")
print(f"F1-score: {f1_rf:.2f}")
print(f"ROC-AUC: {roc_auc_rf:.2f}")


# Лістинг 36. Побудова та оцінка моделі градієнтного бустингу
from sklearn.ensemble import GradientBoostingClassifier
gb_model = GradientBoostingClassifier(n_estimators=200, learning_rate=0.1,
max_depth=5, random_state=42)
gb_model.fit(X_train_bal, y_train_bal)
# Прогнозування
y_pred_gb = gb_model.predict(X_test)
y_proba_gb = gb_model.predict_proba(X_test)[:, 1]
accuracy_gb = accuracy_score(y_test, y_pred_gb)
precision_gb = precision_score(y_test, y_pred_gb)
recall_gb = recall_score(y_test, y_pred_gb)
f1_gb = f1_score(y_test, y_pred_gb)
roc_auc_gb = roc_auc_score(y_test, y_proba_gb)
print("Метрики Gradient Boosting:")
print(f"Accuracy: {accuracy_gb:.2f}")
print(f"Precision: {precision_gb:.2f}")
print(f"Recall: {recall_gb:.2f}")
print(f"F1-score: {f1_gb:.2f}")
print(f"ROC-AUC: {roc_auc_gb:.2f}")

# Лістинг 37. Побудова та оцінка моделі нейронної мережі
```

```
from sklearn.neural_network import MLPClassifier
nn_model = MLPClassifier(hidden_layer_sizes=(64, 32), activation='relu',
solver='adam',
                         max_iter=500, random_state=42)
nn_model.fit(X_train_bal, y_train_bal)
# Прогнозування
y_pred_nn = nn_model.predict(X_test)
y_proba_nn = nn_model.predict_proba(X_test)[:, 1]
accuracy_nn = accuracy_score(y_test, y_pred_nn)
precision_nn = precision_score(y_test, y_pred_nn)
recall_nn = recall_score(y_test, y_pred_nn)
f1_nn = f1_score(y_test, y_pred_nn)
roc_auc_nn = roc_auc_score(y_test, y_proba_nn)
print("Метрики нейронної мережі:")
print(f"Accuracy: {accuracy_nn:.2f}")
print(f"Precision: {precision_nn:.2f}")
print(f"Recall: {recall_nn:.2f}")
print(f"F1-score: {f1_nn:.2f}")
print(f"ROC-AUC: {roc_auc_nn:.2f}")

# Лістинг 38. Побудова та оцінка моделі k-NN
from sklearn.neighbors import KNeighborsClassifier
knn_model = KNeighborsClassifier(n_neighbors=5)
knn_model.fit(X_train, y_train)
# Прогнозування
y_pred_knn = knn_model.predict(X_test)
y_proba_knn = knn_model.predict_proba(X_test)[:, 1]
accuracy_knn = accuracy_score(y_test, y_pred_knn)
precision_knn = precision_score(y_test, y_pred_knn)
recall_knn = recall_score(y_test, y_pred_knn)
f1_knn = f1_score(y_test, y_pred_knn)
roc_auc_knn = roc_auc_score(y_test, y_proba_knn)
print("Метрики k-NN:")
print(f"Accuracy: {accuracy_knn:.2f}")
print(f"Precision: {precision_knn:.2f}")
print(f"Recall: {recall_knn:.2f}")
print(f"F1-score: {f1_knn:.2f}")
print(f"ROC-AUC: {roc_auc_knn:.2f}")

# Лістинг 39. Побудова та оцінка моделі Naive Bayes
from sklearn.naive_bayes import GaussianNB
nb_model = GaussianNB()
nb_model.fit(X_train, y_train)
# Прогнозування
y_pred_nb = nb_model.predict(X_test)
y_proba_nb = nb_model.predict_proba(X_test)[:, 1]
accuracy_nb = accuracy_score(y_test, y_pred_nb)
precision_nb = precision_score(y_test, y_pred_nb)
recall_nb = recall_score(y_test, y_pred_nb)
f1_nb = f1_score(y_test, y_pred_nb)
roc_auc_nb = roc_auc_score(y_test, y_proba_nb)
print("Метрики Naive Bayes:")
print(f"Accuracy: {accuracy_nb:.2f}")
print(f"Precision: {precision_nb:.2f}")
```

```
print(f"Recall: {recall_nb:.2f}")
print(f"F1-score: {f1_nb:.2f}")
print(f"ROC-AUC: {roc_auc_nb:.2f}")

# Лістинг 40. Побудова та оцінка моделі SVM
from sklearn.svm import SVC
svm_model = SVC(probability=True, random_state=42)
svm_model.fit(X_train, y_train)
# Прогнозування
y_pred_svm = svm_model.predict(X_test)
y_proba_svm = svm_model.predict_proba(X_test)[:, 1]
accuracy_svm = accuracy_score(y_test, y_pred_svm)
precision_svm = precision_score(y_test, y_pred_svm)
recall_svm = recall_score(y_test, y_pred_svm)
f1_svm = f1_score(y_test, y_pred_svm)
roc_auc_svm = roc_auc_score(y_test, y_proba_svm)
print("Метрики SVM:")
print(f"Accuracy: {accuracy_svm:.2f}")
print(f"Precision: {precision_svm:.2f}")
print(f"Recall: {recall_svm:.2f}")
print(f"F1-score: {f1_svm:.2f}")
print(f"ROC-AUC: {roc_auc_svm:.2f}")

# Лістинг 41. Таблиця метрик усіх моделей
results = pd.DataFrame({
    "Модель": [
        "Random Forest",
        "Decision Tree",
        "Gradient Boosting",
        "Нейронна мережа",
        "k-NN",
        "Naive Bayes",
        "SVM"
    ],
    "Accuracy": [
        accuracy_rf, accuracy_dt, accuracy_gb, accuracy_nn,
        accuracy_knn, accuracy_nb, accuracy_svm
    ],
    "Precision": [
        precision_rf, precision_dt, precision_gb, precision_nn,
        precision_knn, precision_nb, precision_svm
    ],
    "Recall": [
        recall_rf, recall_dt, recall_gb, recall_nn,
        recall_knn, recall_nb, recall_svm
    ],
    "F1-score": [
        f1_rf, f1_dt, f1_gb, f1_nn,
        f1_knn, f1_nb, f1_svm
    ],
    "ROC-AUC": [
        roc_auc_rf, roc_auc_dt, roc_auc_gb, roc_auc_nn,
        roc_auc_knn, roc_auc_nb, roc_auc_svm
    ]
```

```
})
results.iloc[:, 1:] = results.iloc[:, 1:].round(2)
print(results)

# Лістинг 42. Матриці помилок найкращих моделей
from sklearn.metrics import confusion_matrix
import matplotlib.pyplot as plt
import seaborn as sns
# Список моделей та їхні передбачення
models = {
    "Decision Tree": y_pred_dt,
    "Gradient Boosting": y_pred_gb,
    "Neural Network": y_pred_nn,
    "Naive Bayes": y_pred_nb,
}
# Побудова матриць помилок (4 рядки, 2 стовпці)
fig, axes = plt.subplots(2, 2, figsize=(12, 12))
axes = axes.flatten()
for i, (name, preds) in enumerate(models.items()):
    cm = confusion_matrix(y_test, preds)
    sns.heatmap(
        cm, annot=True, fmt="d", cmap="Blues",
        xticklabels=["No", "Yes"], yticklabels=["No", "Yes"],
        ax=axes[i], cbar=False,
        annot_kws={"size": 20, "weight": "bold"}  # збільшення шрифту
    )
    axes[i].set_title(name, fontsize=20)
    axes[i].set_xlabel("Передбачений клас")
    axes[i].set_ylabel("Фактичний клас")
plt.tight_layout()

# Лістинг 43. Графік важливості ознак
importances = rf_model.feature_importances_
indices = np.argsort(importances)[::-1]
features = X_train.columns
top_n = 20

plt.figure(figsize=(12, 8))
sns.barplot(
    x=importances[indices][:top_n],
    y=features[indices][:top_n],
    palette="viridis"
)
plt.title("10 найважливіших факторів", fontsize=16)
plt.xlabel("Важливість")
plt.ylabel("Ознаки")
# Лістинг 44. Візуалізація дерев рішень
from sklearn.tree import plot_tree
plt.figure(figsize=(18, 10))
plot_tree(
    dt_model,
    feature_names=X_train.columns,
    class_names=["No", "Yes"],
    filled=True,
```

```
    max_depth=3,  # щоб дерево було читабельним
    fontsize=10
)
plt.title("Часткова візуалізація дерева рішень", fontsize=16)

# Лістинг 45. ROC-AUC крива для Decision Tree
from sklearn.metrics import roc_curve, auc
fpr, tpr, thresholds = roc_curve(y_test, y_proba_dt)
roc_auc = auc(fpr, tpr)
plt.figure(figsize=(8, 8))
plt.plot(fpr, tpr, color="darkorange", lw=2, label=f"Decision Tree (AUC =
{roc_auc:.2f})")
plt.plot([0, 1], [0, 1], color="navy", lw=2, linestyle="--")
plt.xlabel("False Positive Rate")
plt.ylabel("True Positive Rate")
plt.title("ROC-AUC крива для моделі Decision Tree")
plt.legend(loc="lower right")

# Лістинг 46. Графік економічного ефекту дерев рішень
data['AnnualIncome'] = data['MonthlyIncome'] * 12
# Фільтруємо тих, хто реально звільнився
leavers = data[data['Attrition'] == 'Yes']
# Загальні витрати на заміну цих працівників
total_replacement_cost = leavers['AnnualIncome'].sum()
# Потенційно передбачені витрати (якби всіх утримати, кого знайшла модель)
saved_cost = total_replacement_cost * recall_dt
# Реалістичні сценарії: утримання лише 5% та 10% працівників
saved_cost_10 = saved_cost * 0.10
print(f"Загальні витрати на заміну звільнених: {total_replacement_cost:,.0f} $")
print(f"Максимально зекономлені кошти (Recall=52%): {saved_cost:,.0f} $")
print(f"Зекономлені кошти при утриманні 10%: {saved_cost_10:,.0f} $")
# Візуалізація
fig, ax = plt.subplots(figsize=(12,10))
bars = ax.bar(
    ["Втрати без моделі",
     "Максимальний ефект",
     "Реалістично (10%)"],
    [total_replacement_cost, saved_cost, saved_cost_10],
    color=["#DC143C", "#2E8B57", "#4682B4", "#6A5ACD"]
)

# Додаємо підписи
for bar in bars:
    height = bar.get_height()
    ax.text(bar.get_x() + bar.get_width()/2, height, f'{height:,.0f} $',
            ha='center', va='bottom', fontsize=20, fontweight='bold')
ax.set_title("Оцінка економічного ефекту від застосування моделі утримання
працівників", fontsize=14)
ax.set_ylabel("Сума, $")a
```