

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДВНЗ «КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА»

Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»

Кафедра математичного моделювання та статистики

Освітньо-професійна програма

Економічна кібернетика

Галузь знань

05 Соціальні та поведінкові науки

Спеціальність

051 Економіка

Форма навчання: очна (денна)

КВАЛІФІКАЦІЙНА БАКАЛАВРСЬКА РОБОТА

на тему «Аналіз ринку праці ІТ-спеціалістів»

(назва теми)

здобувача Подриги Богдана Ігоровича

(ПІБ, підпис)

Науковий керівник: кандидат фізико-математичних наук,
доцент Юнькова О. О.

(науковий ступінь, учене звання, ПІБ)

Робота допущена до захисту перед екзаменаційною комісією
з атестації здобувачів вищої освіти (ЕК)

Завідувач кафедри

канд. фіз-мтмн. наук,

проф. Великоіваненко Г.І.

(підпис)

Київ 2023

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ДВНЗ «КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ ЕКОНОМІЧНИЙ УНІВЕРСИТЕТ
ІМЕНІ ВАДИМА ГЕТЬМАНА»

Навчально-науковий інститут
«Інститут інформаційних технологій в економіці»

Кафедра математичного моделювання та статистики

Освітньо-професійна програма

Галузь знань

Спеціальність

Економічна кібернетика

05 Соціальні та поведінкові науки

051 Економіка

ПОГОДЖЕНО

Керівник проектної групи (гарант)
освітньо-професійної програми

О.М.Притоманова

(підпис) (ініціали, прізвище)

_____ 20__ р.

ЗАТВЕРДЖУЮ

Завідувач кафедри математичного
моделювання та статистики

Г.І.Великоіваненко

(підпис) (ініціали, прізвище)

_____ 20__ р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

здобувачу вищої освіти Подризі Богдану Ігоровичу

(прізвище, ім'я, по батькові)

очної (денної) **форми навчання**

на підготовку кваліфікаційної бакалаврської роботи

на тему **«Аналіз ринку праці ІТ-спеціалістів»**

Тему затверджено наказом ректора Університету від "_____ " _____ 20__ р. № _____

Кваліфікаційна бакалаврська робота виконується на матеріалах

_____ інтернет джерел _____

План кваліфікаційної бакалаврської роботи

Розділ 1	Аналіз функціонування ринку ІТ-послуг в Україні (назва розділу)
Розділ 2	Математичні методи моделювання заробітної плати ІТ спеціалістів (назва розділу)
Розділ 3	Практичне застосування інструментів модельного аналізу до заробітної плати ІТ фахівця (назва розділу)
Об'єкт дослідження:	Ринок праці ІТ спеціалістів
Предмет дослідження:	Моделі і методи економіко-математичного моделювання заробітної плати для виявлення особливостей ринку праці в ІТ-галузі
Мета кваліфікаційної бакалаврської роботи:	Моделювання залежності заробітної плати від впливу різних навичок, компетенцій та досвіду роботи ІТ-спеціаліста, а також аналіз ринку праці ІТ спеціалістів для визначення його стану, актуальних тенденцій та особливостей.

Конкретні завдання, які здобувач повинен виконати для досягнення поставленої мети:

У розділі 1	Дослідити походження та розвиток української ІТ-галузі, вивчити її основні складові та структуру, основні проблеми та недоліки. Проаналізувати вплив пандемії COVID-19 та воєнного конфлікту на ринок праці ІТ спеціалістів в Україні.
У розділі 2	Вивчити теоретичні основи методів математичного моделювання рівня заробітної плати (методи параметричної та непараметричної регресії, методи кластерного аналізу) та їх програмної реалізації
У розділі 3	Застосувати методи параметричної та непараметричної регресії для аналізу факторів, які впливають на рівень заробітної плати ІТ-спеціаліста, виявити структуру середовища працівників ІТ-галузі

**Завдання підготував
науковий керівник**

(підпис)

О.О.Юнькова

(ініціали, прізвище)

« ____ » _____ 20__ р.

**Завдання одержав
здобувач**

(підпис)

Б.І. Подрига

(ініціали, прізвище)

« ____ » _____ 20__ р.

Реферат

Кваліфікаційна бакалаврська робота містить 67 сторінок, 1 таблицю, 35 рисунків, список використаних джерел з 37 найменувань, 1 додаток.

«Аналіз ринку праці ІТ-спеціалістів»

Об'єктом дослідження кваліфікаційної бакалаврської роботи є ринок праці, який включає попит і пропозицію на ІТ спеціалістів в Україні.

Предметом дослідження є моделі і методи економіко-математичного моделювання для виявлення особливостей ринку праці в ІТ-галузі.

Мета і завдання дослідження. Основною метою даної роботи є моделювання залежності заробітної плати від впливу різних навичок, компетенцій та досвіду роботи ІТ-спеціаліста, а також аналіз ринку праці ІТ спеціалістів для визначення його стану, актуальних тенденцій та особливостей.

Відповідно до поставленої мети визначені такі завдання:

- визначити поточний стан та структуру ІТ-ринку України на основі досліджень ІТ-компаній, асоціацій та наукових праць;
- оцінити вплив пандемії та війни на український ІТ-ринок;
- розглянути інструменти та засоби моделювання залежності заробітної плати ІТ спеціаліста від його фахових компетентностей та характеристик, та особливості формування заробітної плати окремих груп фахівців;
- побудувати модель покрокової регресії та оцінити вплив різних навичок, досвіду роботи, освіти та розміру компанії на заробітну плату ІТ-спеціаліста, перевірити якісні характеристики моделі;
- провести групування ІТ спеціалістів засобами кластерного аналізу та охарактеризувати ці групи;
- здійснити прогнозування за допомогою кращих моделей та оцінити результати

Теоретична, методична та практична значущість отриманих результатів. Теоретична та методична значущість даної роботи полягає в поєднанні різних методів математичного моделювання заробітної плати працівників української ІТ-галузі, вивченні структури середовища її фахівців.

Практична значущість даної роботи полягає у визначенні впливу різних факторів, таких як досвід роботи, навички, освіта та компанія, в якій працює ІТ-фахівець, на заробітну плату ІТ-спеціаліста. Це допоможе рекрутинговим агенціям краще зрозуміти ситуацію на ринку праці, а ІТ-спеціалістам – над якими навичками працювати задля збільшення заробітної плати.

Рік виконання кваліфікаційної бакалаврської роботи – 2023.

Рік захисту роботи – 2023.

Ключові слова: ІТ-ринок, ІТ-спеціаліст, заробітна плата, лінійна регресійна модель, моделювання, кластерний аналіз, прогнозування

В і д г у к
про кваліфікаційну бакалаврську роботу
здобувача навчально-наукового інституту ІТЕ
освітньо-професійної програми «Економічна кібернетика»
Подриги Богдана Івановича
на тему **«Аналіз ринку праці ІТ-спеціалістів»**

1. Актуальність теми: У зв'язку з непростотою економічною ситуацією в Україні, однією з найактуальніших проблем стає тема заробітної плати, як основного джерела забезпечення людини необхідними благами. Заробітна плата не є стабільною, вона має тенденцію до диференціації по різних видах діяльності, умовам праці, по регіонах та по самим індивідуумам. Звідси постає проблема визначення того, на що опиратись і на що можна розраховувати працівникам і роботодавцям у найближчий час.
2. Позитивні риси кваліфікаційної роботи: Проаналізовано розвиток українського ІТ ринку, описано його структуру та особливості функціонування в умовах пандемії та війни. Досліджено вплив різних факторів на заробітну плату ІТ спеціаліста методами параметричної та непараметричної регресії і кластерного аналізу.
3. Наявність самостійних розробок автора: Дослідження виконується на реальних даних, підібраних здобувачем. Методи математичного моделювання та статистичного аналізу реалізовані в середовищі RStudio, отримані результати інтерпретовані на достатньому науковому рівні. Це засвідчує високий рівень володіння сучасними аналітичними методами та інформаційними технологіями.
4. Цінність теоретичних висновків та практичних рекомендацій: запропоновані в роботі механізми для моделювання заробітної плати можуть бути застосовані у будь-якій галузі чи регіоні. Отримані в роботі висновки дають змогу працівникам визначати напрямки підвищення своєї кваліфікації, а роботодавцям оптимізувати пошук спеціалістів.
5. Наявність недоліків: В роботі можна було б врахувати більше аспектів формування заробітної плати ІТ спеціалістів. Однак, ці побажання не впливають на загальну якість виконаної роботи.
6. Загальна оцінка кваліфікаційної бакалаврської роботи та її допущення до захисту перед ЕК: Вважаю, що кваліфікаційна бакалаврська робота Подриги Б.І. «Аналіз ринку праці ІТ-спеціалістів» цілком відповідає вимогам, які висуваються до даного типу робіт. Робота виконана на достатньо високому фаховому рівні, може бути рекомендована до захисту перед ЕК і заслуговує позитивної оцінки.

Науковий керівник:
професор кафедри
математичного моделювання та статистики
канд.ф.-м.наук, доцент

Юнькова О.О.

“ ____ ” _____ 20__ р.

ЗМІСТ

ВСТУП	7
РОЗДІЛ 1 АНАЛІЗ ТА ПРИНЦИПИ ФУНКЦІОНУВАННЯ РИНКУ ІТ ПОСЛУГ В УКРАЇНІ	10
1.1 Структура ІТ ринку в Україні.....	10
1.2 ІТ галузь під час пандемії та війни	16
1.3 Головні проблеми та недоліки ринку ІТ в Україні	21
РОЗДІЛ 2 МАТЕМАТИЧНІ МЕТОДИ МОДЕЛЮВАННЯ ЗАРОБІТНОЇ ПЛАТИ ІТ СПЕЦІАЛІСТІВ.....	25
2.1 Лінійна регресійна модель залежності заробітної плати від впливових чинників.....	25
2.2 Методи непараметричної регресії для прогнозування заробітної плати	27
2.3 Кластерний аналіз та метод k-середніх	29
РОЗДІЛ 3 ПРАКТИЧНЕ ЗАСТОСУВАННЯ МОДЕЛЬНОГО ІНСТРУМЕНТАЛЬНОГО АНАЛІЗУ ДО ЗАРОБІТНОЇ ПЛАТИ ІТ ФАХІВЦЯ	35
3.1 Модель прямої покрокової регресії	35
3.2 Вплив якісних показників на заробітну плату.....	38
3.3 Порівняння результатів прогнозування моделей.....	48
3.4 Кластерний аналіз	49
ВИСНОВКИ.....	55
СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ.....	57

ВСТУП

В умовах технологічного прогресу, зростання попиту на цифрові рішення та активного застосування інформаційних технологій у кожній сфері діяльності людини, ринок ІТ розвивається дуже стрімко. ІТ є однією з провідних галузей української економіки, яка навіть в умовах пандемії та воєнного стану не втратила своїх позицій, а навіть зросла та позитивно впливає на економіку держави.

Незважаючи на те, що ІТ-галузь здебільшого орієнтована на експорт, існує багато програмних продуктів, розроблених українськими фахівцями, які успішно працюють в різних галузях економіки країни і дозволяють автоматизувати виробничі процеси, підвищити продуктивність, знизити витрати та загалом збільшити ефективність бізнесу. Як показало дослідження Асоціації ІТ Ukraine, проведене у жовтні-листопаді 2022 року, 21,7% підприємств мають штатних ІТ-фахівців, а ще 14,6% підприємств залучають позаштатні ресурси для виконання функцій ІКТ (інформаційно-комунікаційних технологій).

Згідно офіційної статистики частка експорту комп'ютерних послуг становить 37,8%, тобто більше третини від усіх послуг, які експортує Україна на зовнішній ринок, при тому, що в ІТ працює лише 2% від усіх зайнятих осіб в країні.

Висока прибутковість даної галузі та високі заробітні плати її фахівців є предметом інтересу багатьох дослідників. Але в умовах війни, коли економіка країни зазнає значних змін, зокрема через відтік кадрів, дослідження ринку праці ІТ набуває особливої актуальності.

Дослідження заробітної плати в різних сферах української економіки проводили такі вітчизняні вчені, як: А. Мельник, В. Прима, Н. Жмурко, Н. Карпенко, Т. Філіпенко, Т. Мельник, Є. Завгородня, Х. Мацелюх. Вивченням проблем ІТ-ринку займалась Г. Шведова, а моделюванням заробітної плати займались В. Молчан та Т. Маринич.

Метою даної роботи є моделювання залежності заробітної плати від впливу різних навичок, компетенцій та досвіду роботи ІТ-спеціаліста, а також аналіз ринку праці ІТ спеціалістів для визначення його стану, актуальних тенденцій та особливостей.

Для досягнення поставленої мети потрібно виконати такі завдання:

1. Визначити поточний стан та структуру ІТ-ринку України на основі досліджень ІТ-компаній, асоціацій та наукових праць.
2. Оцінити вплив пандемії та війни на український ІТ-ринок.
3. Розглянути інструменти та засоби моделювання залежності заробітної плати ІТ спеціаліста від його фахових компетентностей та характеристик, та особливості формування заробітної плати окремих груп фахівців.
4. Побудувати модель покрокової регресії та оцінити вплив різних навичок, досвіду роботи, освіти та розміру компанії на заробітну плату ІТ-спеціаліста, перевірити якісні характеристики моделі.
5. Провести групування ІТ спеціалістів засобами кластерного аналізу та охарактеризувати ці групи.
6. Здійснити прогнозування за допомогою кращих моделей та оцінити результати.

Об'єктом дослідження є ринок праці, який включає попит і пропозицію на ІТ спеціалістів в Україні.

Предмет дослідження – моделі і методи економіко-математичного моделювання для виявлення особливостей ринку праці в ІТ-галузі.

У кваліфікаційній бакалаврській роботі були використані статистичні методи (зведення, групування, узагальнюючі характеристики), методи регресійного і кластерного аналізу, а також випадковий ліс для прогнозування.

Інформаційна база дослідження включає звіти ІТ-асоціацій та компаній, наукові праці, які стосуються структури, проблематики та принципів функціонування ІТ-ринку, інтернет-джерела щодо методів моделювання та

прогнозування, які можуть використовуватись для вивчення впливу різних факторів на заробітну плату ІТ-спеціаліста.

Структура роботи включає вступ, три розділи, висновки, список використаної літератури та додатки.

РОЗДІЛ 1

АНАЛІЗ ТА ПРИНЦИПИ ФУНКЦІОНУВАННЯ РИНКУ ІТ ПОСЛУГ В УКРАЇНІ

1.1 Структура ІТ ринку в Україні

ІТ-ринок в Україні - це динамічна та швидко зростаюча галузь, яка відіграє важливу роль у розвитку української економіки. Він включає в себе широкий спектр діяльності, від розробки програмного забезпечення та веб-розробки до аналізу даних, штучного інтелекту, хмарних технологій та кібербезпеки.

Однією з основних переваг ІТ-ринку в Україні є значна кількість висококваліфікованих ІТ-спеціалістів. Країна має велику кількість вищих навчальних закладів, які готують фахівців у сфері інформаційних технологій. Багато українських програмістів та розробників зарекомендували себе на міжнародному ринку завдяки своєму технічному ноу-хау та високому рівню професіоналізму.

ІТ-компанії в Україні активно співпрацюють з клієнтами з усього світу. Вони надають різноманітні ІТ-послуги, включаючи розробку програмного забезпечення на замовлення, тестування, підтримку та інтеграцію систем. Українські ІТ-компанії часто володіють високим рівнем експертизи в конкретних сферах, таких як фінанси, медицина, логістика, електронна комерція та багато інших.

За останні 10 років ІТ-сектор став значним гравцем в економіці України. З розвитком ІТ-освіти все більше і більше людей працюють у цій галузі, поповнюючи бюджет України. Зростав і розвиток ІТ-кластеру. З'являлись такі хаби як Lyft99 та Unit.City, де розвивались та зростали стартапи та продуктові компанії. Популяризувались конференції та мітапи. В очах пересічного українця ІТ-сектор

виріс із гківського захоплення студентів-ентузіастів у велич та красу. Наразі багато людей хочуть бути частиною цього світу, а ще більше – працювати та створювати продукти, знаючи, що за цим – майбутнє.

Зкладами ІТ-освіти в Україні є університети, коледжі, курси та ІТ-школи. У 2021 р. майже 1 200 закладів вищої освіти навчали студентів в Україні. Більш ніж 300 з них навчали ІТ-фахівців різних рівнів та спеціальностей.

З 2017 року прослідковується стрімке зростання кількості вступників до вищих навчальних закладів. За 5 років кількість вступників до бакалаврату збільшилась майже на 50% з 20,3 тис. до 30,1 тис. осіб. Однак офіційна освіта зазвичай не може покрити всіх потреб ІТ-галузі, тому більша частина фахівців потребує додаткової освіти. В Україні вона представлена ІТ-школами та курсами (рис. 1.1).

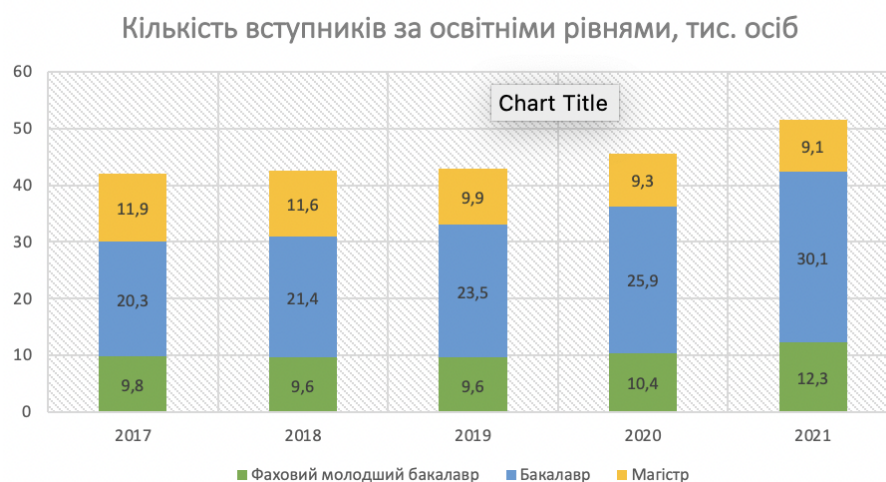


Рисунок 1.1 – Кількість вступників за освітніми рівнями, тис. осіб

Джерело: розроблено автором на основі [1]

Активний розвиток ІТ-галузі зумовлює зростання кількості ІТ-фахівців. З 2015 р. їх чисельність зросла на 116% з 43,4 тис. активних найманих працівників та 90,1 тис. ФОПів до 63,1 тис. та 243,9 тис. відповідно, здебільшого за рахунок зростання активних ФОПів, адже така взаємодія для компаній є зручнішою (рис. 1.2).



Рисунок 1.2 – Чисельність ІТ-фахівців в Україні, тис. осіб

Джерело: розроблено автором на основі [1]

Згідно дослідженнями Ukrainian Tech Ecosystem Overview [2] в Україні станом на 2022 рік працює 2234 ІТ-компанії. З них 1617 продуктових компанії та 535 сервісних компанії (рис. 1.3).



Рисунок 1.3 – Кількість компаній різного типу на ІТ-ринку України [2]

Офіси компаній переважно знаходяться в Києві: 1406 продуктових та 348 сервісних компаній. Друге місце у рейтингу посідає Львів з 143 офісами продуктових та 146 сервісних компаній. Третє місце належить Харкову (117 та 109 офісів відповідно) (рис. 1.4).

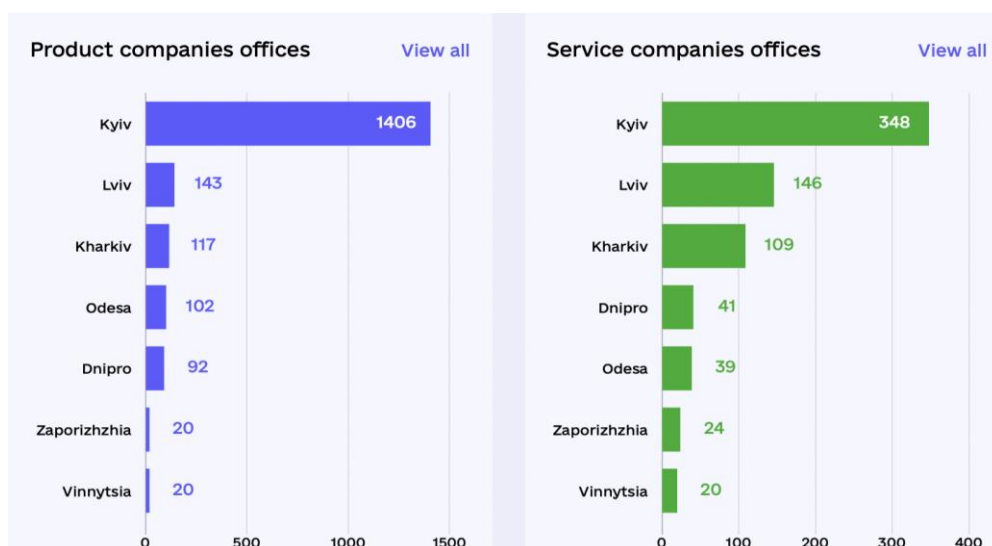


Рисунок 1.4 – Кількість офісів ІТ-компаній в Україні за містами [2]

312 компаній працюють у сфері Martech & Media, 291 – у сфері Business productivity software, 195 – у сфері Fintech & Insurtech, 141 – Hardware & IoT, 130 – E-commerce & Retail, 125 – ESports & Gaming, 120 – Edtech, 62 – Costumer products, 58 – AI. Більшість сервісних компаній мають від 21 до 50 найманих працівників, друге місце посідають компанії із 101 – 250 найманими працівниками і всього лиш 15 компаній налічують у штаті більше 1000 осіб (рис 1.5).

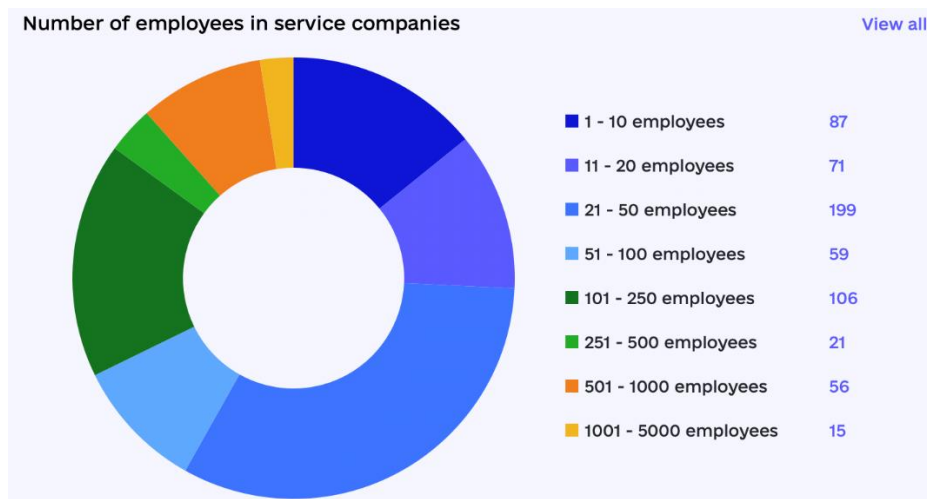


Рисунок 1.5 – Кількість сервісних компаній щодо чисельності працівників[2]

Загалом компанії можна поділити на 3 види: продуктові, сервісні, які поділяються на аутсорсинг, аутстафінг та консалтингові компанії.

Продуктові компанії займаються розробкою власного додатку, веб-сайту або ж сервісу. Це називається продуктом. Такі компанії займаються не тільки розробкою свого продукту, а також маркетинговою та PR частиною. Їхніми клієнтами є користувачі продукту. Найвідомішими українськими компаніями цього виду є Grammarly, GitLab, TemplateMonster, Jooble, Depositphotos та Jooble.

Сервісні IT-компанії працюють на основі прийому замовлень від інших суб'єктів та виконання IT-задач згідно з умовами договору. Вони зазвичай спеціалізуються на обслуговуванні закордонних клієнтів, які потребують реалізації конкретних IT-проектів, зокрема розробки програмного забезпечення. Їхні послуги можуть включати також консультування та інші IT-послуги.

Сервісні IT-компанії можуть використовувати різні моделі роботи такі як: аутсорсинг, аутстафінг та надання консалтингових послуг. Вони пропонують свої послуги з метою задоволення потреб клієнтів та отримання відповідної компенсації [3].

Різниця між аутсорс та аутстаф компаніями полягає в тому, що перші надають клієнтам команди для розробки, а другі – окремих спеціалістів. Найвідомішими аутсорс компаніями є: EPAM Systems, SoftServe, GlobalLogic, Ciklum, Intellias. Щодо аутстаф компаній – сміливо можна назвати SvitlaSystems, N-iX, Daxx, Cogniance, Beetroot.

Щодо консалтингових компаній, можна сказати, що такі компанії сприяють впровадженню готових продуктів програмного забезпечення, які були розроблені іншими IT-фірмами або самими замовниками. На відміну від інших IT-сервісних компаній, вони не займаються самостійною розробкою програмного продукту для інших і не надають послуг з розробки. Вони просто допомагають звичайним підприємствам забезпечити стабільну роботу або розвиток шляхом використання придбаного ними комп'ютерного обладнання та/або програмного забезпечення. Основне завдання цих компаній полягає в налаштуванні стабільної роботи різних IT-компонентів. Проте, насправді, цей процес є складнішим. Часто аутстафінгові

компанії пропонують свої послуги в якості консультантів з питань інформатизації [4].

Якщо ж розглядати плюси та мінуси роботи в кожній із компаній, можна сказати таке. У продуктових компаніях працівники повністю занурені у свій продукт та працюють завжди в одній команді. Робота в таких компаніях супроводжується самовдосконаленням та розвитком разом з продуктом. В таких компаніях вітаються люди, які приносять творчість та нові ідеї. Недоліками є однотипна робота та можливість втратити роботу, якщо продукт стане нікому не потрібним. В сервісних компаніях перевагами для працівника є різноманітність сфер, проєктів, продуктів та завдань, які стоять перед працівником. Через це розширюється коло знайомств та, зазвичай, через це різноманіття, швидко ростуть навички робітника. Недоліками є орієнтування компанії не на продукт, а на бажання замовника, робота на швидкість та надмірна залежність від замовника. Фахівці у консалтингових компаніях зазвичай мають великі зарплатні, оскільки це тривалий енерговитратний процес, що вимагає повного занурення у справи та роботу компанії. Недоліками у цьому разі є підвищений вплив стресу та часті відрядження, якщо фахівець працює не онлайн, що трапляється дуже часто.

Протягом останніх десятиліть, українські ІТ компанії здобули визнання як на внутрішньому, так і на міжнародному ринку послуг. Вони стали справжніми гравцями на ІТ ринку, надаючи широкий спектр послуг, таких як розробка програмного забезпечення, хмарні технології, сервісний дизайн, кібербезпека, робототехніка та багато іншого.

Українські компанії вразили світ не лише своєю розмаїтістю послуг, а й великими досягненнями на міжнародному ринку. Багато з них входять до списку найбільших ІТ компаній світу. Це свідчить про те, що українські фахівці мають високий рівень компетенції та здатні конкурувати з кращими гравцями глобального ІТ сектора.

Серед найбільших ІТ-компаній України можна виділити такі компанії як SoftServe, EPAM Systems та GlobalLogic.

1.2 ІТ галузь під час пандемії та війни

Випробуванням для ІТ-галузі стала пандемія у 2020-2021р. Під час карантину багато компаній почали працювати онлайн і це стало викликом для багатьох бізнес-процесів. Більшість фахівців всі ці роки працювали офлайн в офісах, де були налагоджені комунікації, відповідальність та контроль. Для людей офісом став їхній власний дім і тут їх спіткали найбільші проблеми, такі як самоізоляція, зміна обстановки, домашня рутина, недостатня самоорганізованість. Згодом абсолютна більшість фахівців адаптувалась до цих змін, які в кінцевому рахунку виявилися навіть на краще, бо почались значні зміни на ринку праці. В онлайні наймані працівники не прив'язані до свого місця розташування, тому перед ними відкрилися нові можливості пошуку і отримання роботи по всьому світу.

Віддалена робота дозволила компаніям бути орієнтованими на співробітників, а не на офіс [5]. Квадратні метри, колір стін і зручність офісних крісел раптово втратили сенс. ІТ-компанії зосередили велику увагу на елементах корпоративної атмосфери та культури. У центрі уваги стала людина – співробітник, який є частиною компанії, незалежно від того, де він перебуває. Під час пандемії змінилися процеси найму та адаптації та активно впроваджувалися ініціативи, щоб зробити співробітників більш продуктивними в умовах невизначеності, стали піклуватися про їх психічне здоров'я та благополуччя, а співробітники охоче розвивалися та вчилися, навіть якщо вони мали надзвичайно напружений графік онлайн.

У першому півріччі 2020 український ІТ-ринок зазнав скорочення, але деякі сектори, зокрема e-commerce, розробка ігор, онлайн-навчання, івенти, автоматизація бізнесу у сферах доставок та продуктового ритейлу життєво процвітали.

Найбільше постраждали торгівля комп'ютерами та гаджетами, а також послуги з системної інтеграції. Купівельна активність в цій підгалузі зменшилась

через падіння доходів та відкладення покупок. Послуги системної інтеграції також постраждали через зменшення бюджетів урядових та корпоративних секторів на потреби в ІТ.

Однак, розвиток онлайн-сервісів у період карантину збільшив популярність підписок на різноманітні онлайн-послуги: новини, музика, розваги, доставка, навчання. В корпоративному секторі зросла потреба у OTT-сервісах, відеоконференціях та віддаленій роботі. Значно зросла також інтернет-торгівля [6].

Найважливіше, що 2020 рік прискорив процес цифрової трансформації бізнесу та суспільства загалом. Навіть при падінні доходів витрати на ІТ-потреби в державних установах, корпораціях, МСБ та приватних осіб зросли.

За 2021 рік, згідно з дослідженнями Асоціації «ІТ Україна» ІТ галузь зросла на 36%, досягнувши результату в \$6,8 млрд. Кількість спеціалістів збільшилась з 244 тис. до 285 тис. 2021 рік став роком кандидатів. Ринок кандидатів - це термін, який використовується для опису ринкової ситуації, коли попит на кваліфікованих працівників перевищує їх пропозицію. Це означає, що компанії активно шукають працівників з певними навичками та досвідом, але знаходять обмежену кількість кандидатів. Конкуренція серед роботодавців відіграє важливу роль на ринку кандидатів, що змушує їх конкурувати за найкращих працівників. Кандидати мають більше варіантів роботи та можуть вибирати між пропозиціями від різних компаній. Це призводить до підвищення заробітної плати, кращих умов праці та інших переваг для кандидатів.

Саме така ситуація складалась в Україні в 2021 році. Через це багато людей хотіли увійти в ІТ та ІТ курси набрали шалену популярність. За різними оцінками кількість випускників таких курсів складала від 10 до 12 тисяч осіб [8].

З початком війни ІТ-компанії виявилися дійсно підготовленими до евакуації співробітників, організували її в максимально короткі терміни. Всі забули про бюрократію, списки про евакуацію складали максимально швидко. Евакуювали не тільки співробітників, а й їх родичів та домашніх улюбленців і

навіть друзів. По прибуттю до місць евакуації, компанії оплачували харчування та житло.

53% ІТ компаній здійснили часткову релокацію, 17,5% - повну, а 29,2% компаній не релокувались зовсім. З них 17,5% здійснили релокацію в Україні, 49,6% як в Україні так і за її межами, а тільки 3,6% - релокувались повністю за межі України (рис 1.6).

Релокація ІТ-компаній через війну

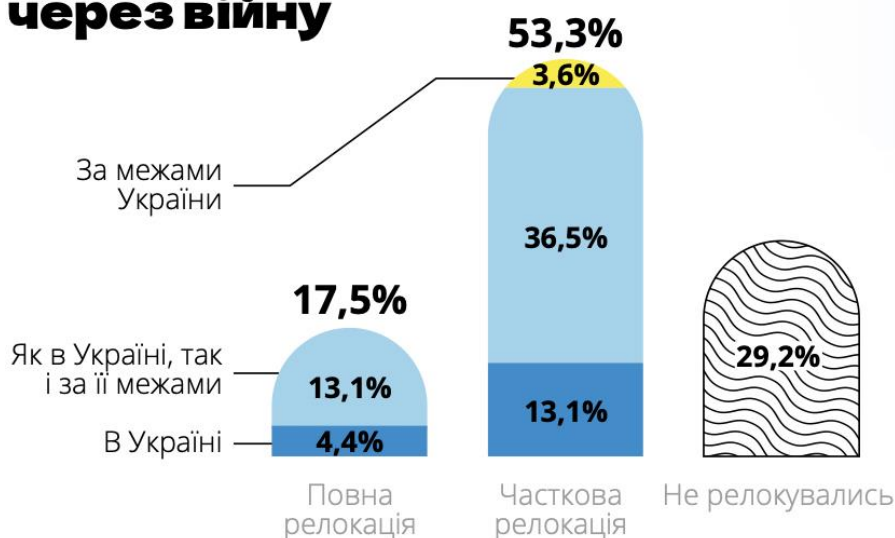


Рисунок 1.6 – Релокація ІТ-компаній через війну [1]

До ТОП-10 країн, до яких релокувались ІТ-компанії, входять Польща, Німеччина, США та інші. Більше 40% ІТ компаній, які релокувались, вибрали Польщу через податкову систему, малу відстань між державами та серйозну підтримку біженців і релокації компаній (рис 1.7).

ТОП-10 країн, до яких релокувались ІТ-компанії

Частка компаній, що обрала відповідний варіант, %

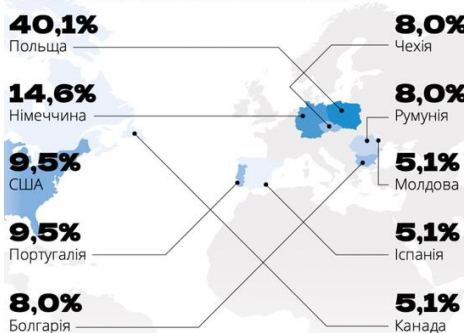


Рисунок 1.7 – Топ-10 країн, до яких релокують ІТ-компанії [1]

Після евакуації співробітники ІТ-компаній почали піднімати питання щодо офісів в Росії та сплати там податків, що є прямим фінансуванням війни, часто відмовлялись навіть працювати в командах, де членами яких були громадяни країни агресора [11].

Коли почалась війна, клієнти українських ІТ-компаній були налаштовані скептично та песимістично щодо української розробки. Однак як тільки російські війська були відтиснуті з-під Києва, Чернігова та Харкова, ситуація почала стабілізуватися, західні замовники повернулися до українських розробників.

Бізнес в перші місяці війни повністю відволікся від своїх основних задач, а почав займатись питаннями свого збереження та виживання, при цьому продовжуючи платити заробітні плати та наймаючи нових співробітників, купуючи нову техніку та навіть займаючись освітніми програмами.

Війна підірвала український ІТ ринок та протягом весни 2022 року мало хто взагалі задумувався про найм кандидатів [12]. Кількість вакансій за різними напрямками зменшилась практично вдвоє, а кількість вакансій, згідно аналітиці DOU.ua, зменшилась у 6-7 разів. На порталі jobs.ua за перший тиждень війни було опубліковано 370 вакансій, у той час, як за тиждень до вторгнення їх було 2,5 тисячі [13].

Наступним викликом для української ІТ індустрії стали масовані обстріли енергосистеми України, які здійснювались протягом осені-зими 2022 року. Багато людей залишались днями без світла та зв'язку, що унеможливлювало роботу, тому люди та бізнеси почали закуповувати старлінки та генератори для безперебійного інтернету та світла. Звісно, не всі мали таку змогу, тому це стало ще одним випробуванням для ІТ-галузі.

Абсолютна більшість ІТ-компаній придбали генератори та використовували Starlink. Більше половини - закуповували паливо та створювали системи інформування. Близько 40% диверсифіковували інтернет-провайдерів та закуповували павербанки. Близько 30% перевели дані «у хмару» та облаштували офіси з можливістю проживання команди і всього лиш 16% оплачували коворкінг.

Всі ці заходи допомогли компаніям підтримувати бізнес-процеси та не збанкрутувати (рис 1.8).



Рисунок 1.8 – Топ-10 країн, до яких релокують ІТ-компанії

Джерело: розроблено автором на основі [1]

Проте, за підсумками 2022 року, згідно дослідження асоціації IT Ukraine [1], ІТ стала однією з небагатьох галузей української економіки, яка змогла не лише зберегти обсяги експорту послуг, а й збільшити їх. Обсяг експорту ІТ-послуг зріс на 9,9% порівняно з попереднім роком та сягнув відмітки у 5 млрд дол. США.

У лютому 2023 спільнота DOU.ua провела велике опитування ІТ-фахівців та дослідила ринок праці [14]. За результатами опитування виявилось, що 84% людей працюють, а 13% - без роботи та активно її шукають (рис 1.9).



Рисунок 1.9 – Ситуація із зайнятістю ІТ-спеціалістів [14]

40% спеціалістів шукали нову компанію чи проєкт у 2022 році, що говорить про закриття багатьох проєктів та скорочення штату працівників в компаніях, яке сталося якраз навесні 2022 року (рис. 1.10).

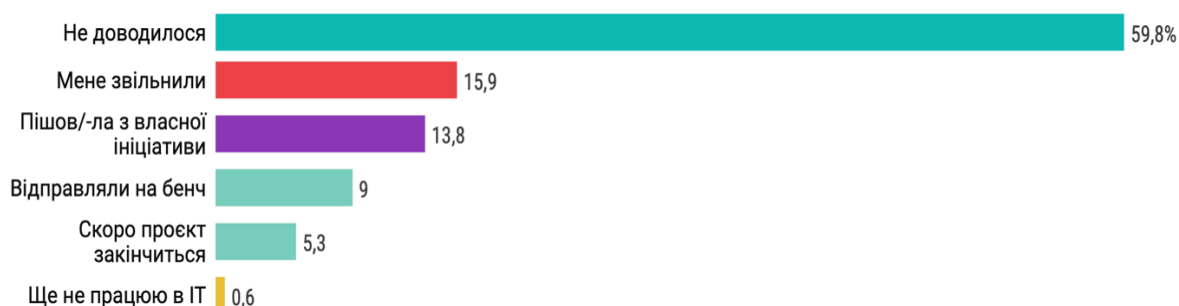


Рисунок 1.10 – Пошук роботи ІТ-спеціалістами в 2022 році[14]

Понад 50% айтівців хвилюються через можливу втрату роботи та вважають, що їм було б важко або дуже важко знайти роботу. Також 38% спеціалістів в Україні й 48% за кордоном вважають, що у 2023 році українське ІТ чекає зменшення кількості робочих місць.

Найкращими новинами для спеціалістів є те, що заробітні плати майже не змінилися. Середньостатистичний Junior може розраховувати на 500-1500\$, Middle – 2000-3000\$, Senior - 3000-10000\$.

1.3 Головні проблеми та недоліки ринку ІТ в Україні

Незважаючи на те, що якість української розробки відома у всьому світі, ринок ІТ в Україні має свої недоліки та проблеми.

Першою проблемою є низькі заробітні плати, порівняно із США та Європою. Причиною цього є недостатня розвиненість економіки України, порівняно з іншими провідними державами. Низькі заробітні плати можуть слугувати причиною для втрати талановитих фахівців, які мають можливість виїзду закордон заради кращого життя.

Карпенко Н. В. вважає, що для максимального використання наявного ІТ-потенціалу Україні потрібно запровадити прозору податкову реформу, забезпечити надійний і дієвий захист інтелектуальної власності та налагодити тісний

взаємозв'язок підприємств ІТ-галузі із закладами освіти та урядовими структурами [16].

Нестабільна політична та економічна ситуація є великою проблемою для українського ІТ ринку. Майдан 2004 року, другий майдан 2013-2014 років, з 2014 року - війна на сході країни, а з 2022 – повномасштабне вторгнення. Всі ці події негативно впливають на зростання ІТ ринку та зупиняють інвесторів і провідні міжнародні компанії щодо співпраці з Україною.

Інфраструктура в Україні часто не відповідає потребам сучасної ІТ-індустрії. Наприклад, нестабільність електропостачання може призвести до порушення роботи комп'ютерів та інших пристроїв, що негативно позначається на продуктивності компаній. Крім того, якість Інтернету не завжди висока, що може обмежувати швидкість і доступність послуги, особливо в невеликих містах та селах.

Северук І. М. та Філіпенко Т.Г. зосереджують свою увагу на проблематиці інфраструктури ІТ в Україні. Науковці пропонують розширити можливості доступу населення до комп'ютерних та інформаційних технологій, розповсюджувати ІТ-послуги та створення загальнодоступних електронних ресурсів за рахунок держави і бізнесу [17].

Хоча Україна має значний потенціал в ІТ-секторі та має великий резерв талановитих професіоналів, все ще відчувається дефіцит кваліфікованих кадрів у певних сферах знань. Наприклад, штучний інтелект, машинне навчання, кібербезпека та інші нові технології вимагають спеціалістів із певними навичками та досвідом.

В свою чергу, Кондіус І.С. наголошує на кадровому питанні та говорить про те, що освітні системи є застарілими. Головною причиною є те, що отримуючи дозвіл на право навчання за певними спеціальностями, вуз зобов'язаний узгодити з міністерством навчальну програму, в якій директивно прописано безліч дисциплін аж до окремих тем, які, насправді, дуже віддалено співвідносяться з потребами реальних компаній. А компанії, в свою чергу створюють власні навчальні

підрозділи, навчання в яких проводиться безкоштовно, а найбільш здатним випускникам надається можливість відразу ж отримати роботу [18].

Також про нестачу спеціалістів згадують Н.О. Бринза, А.А. Гаврилова. Ця нестача пов'язана з тим, що одночасно існують проблеми як з підготовки конкурентоспроможних спеціалістів профільними навчальними закладами, так й “вимивання” кваліфікованих кадрів не тільки з менш благополучних регіонів в вітчизняні IT-центри, а й за межі країни.

На думку Вікулової А.О. та Савчук В.В. подальший розвиток ринку IT-послуг в Україні можливий за умови гарантування якісної освіти та підготовки конкурентоспроможних IT-фахівців, сприяння розвитку внутрішнього ринку, а також посилення законотворчих процесів у сфері інтелектуальної власності та захисту прав інвесторів [19].

IT-ринок в Україні конкурує з іншими країнами, такими як Індія, Польща та Вірменія, у сфері IT-аутсорсингу та розробки програмного забезпечення. Це створює велику конкуренцію між підприємствами та може знизити прибутковість. Компанії повинні знайти свої конкурентні переваги, такі як: висока якість послуг, ефективність та інновації, щоб вижити на цьому ринку.

Працюючи з клієнтами та партнерами з інших країн, українські компанії зіштовхуються з культурними відмінностями та мовними бар'єрами. Розбіжності в розумінні ділової практики, стилів роботи та культурних нюансів можуть призвести до непорозумінь і ускладнити ефективне спілкування. Компанії повинні інвестувати в міжкультурну освіту та розвивати навички міжнародної комунікації для успішної співпраці з партнерами з інших країн.

Також однією з основних проблем є висока ймовірність економічного дисбалансу між регіонами України, що може виникнути в результаті масової міграції IT-фахівців та IT-кластерів зі сходу країни на захід, у зв'язку зі значним руйнуванням економічної та соціальної інфраструктури багатьох міст через воєнний конфлікт.

Урядова увага до відновлення та розвитку інфраструктури східних регіонів України буде одним з найважливіших напрямків післявоєнної відбудови країни. Цей процес буде вимагати значних фінансових ресурсів, що будуть виділятися з державного бюджету. Формування такого бюджету вимагатиме введення податків та сприятиме залученню компаній до роботи в цих регіонах. Варто зауважити, що на сході країни нині майже відсутні компанії, тому цей процес буде включати створення сприятливих умов для приваблення бізнесу та інвестицій [20].

Ці недоліки та обмеження не є неминучими, але їх потрібно усунути. Уряди, підприємства та навчальні заклади можуть працювати разом, щоб розвивати інфраструктуру, покращувати якість освіти та навчання, просувати інновації та покращувати міжнародне спілкування, щоб покращити ситуацію.

РОЗДІЛ 2

МАТЕМАТИЧНІ МЕТОДИ МОДЕЛЮВАННЯ ЗАРОБІТНОЇ ПЛАТИ ІТ СПЕЦІАЛІСТІВ

2.1 Лінійна регресійна модель залежності заробітної плати від впливових чинників

Заробітна плата фахівців ІТ сектору формується під впливом багатьох факторів. Цей вплив не є детермінованим, але є статистичним. Різновидом статистичного зв'язку є зв'язок кореляційний. Він полягає в тому, що при змінюванні одного показника змінюється середнє значення іншого. Для відображення кореляційних зв'язків застосовуються регресійні рівняння.

Лінійна регресія - це лінійна модель, тобто модель, яка відображає пропорційний вплив одного чи кількох чинників на деякий результативний показник. Лінійна залежність між вхідною змінною (x) і вихідною змінною (y) називається простою (парною) лінійною регресією. Якщо вхідних змінних декілька, модель називається множинною лінійною регресією і має вигляд [21]:

$$y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u \quad (2.1)$$

y – залежна змінна,

$(x_1, x_2, \dots, x_k)$ – вектор незалежних змінних;

$(\beta_0, \beta_1, \dots, \beta_k)$ – вектор параметрів, які потрібно обчислити на підставі статистичних даних;

u – випадкова похибка, розподіл якої в загальному випадку залежить від незалежних змінних, але математичне сподівання якої рівне нулю.

Побудова лінійної регресійної моделі полягає у виборі параметрів $\beta_0, \beta_1, \beta_2, \dots, \beta_k$ (оцінювання) в такий спосіб, щоб мінімізувати суму квадратів

залишків, які відображають розбіжність між спостереженими значеннями відгуку і тими, що обчислюються за моделлю.

Однією з узагальнюючих характеристик моделі є коефіцієнт детермінації R^2 :

$$R^2 = 1 - \frac{SSE}{SST} \quad (2.2)$$

R^2 – коефіцієнт детермінації,

SSE – сума квадратів залишків регресії,

SST – сума квадратів відхилень точок даних від середнього значення

R^2 відображає відношення варіації залежної змінної, яку пояснює модель, до загальної варіації залежної змінної. Він може набувати значення від 0 до 1, де 0 означає, що модель не пояснює жодної варіації, а 1 вказує на повну відповідність моделі до даних.

Близькі до 1 значення R^2 свідчать про кращу адекватність моделі до даних. Проте, R^2 не вказує на причинно-наслідкові зв'язки або якість прогнозування моделі. Він може бути підвищений шляхом додавання додаткових змінних до моделі, навіть якщо ці змінні не мають сильного впливу на залежну змінну.

Т-тест (t-test) використовується як метод перевірки регресійної моделі в частині оцінювання статистичної значущості коефіцієнтів регресії. Він допомагає визначити, чи є коефіцієнти статистично відмінними від нуля. Якщо значення t-статистики є значущим (тобто вище певного порогового значення), то можна прийняти альтернативну гіпотезу, що коефіцієнт не є нульовим.

При оцінюванні значущості кожного коефіцієнта регресії, t-тест використовується для обчислення значення t-статистики. Ця статистика обчислюється діленням оцінки коефіцієнта на його стандартну помилку. Значення t-статистики порівнюється з критичним значенням, яке залежить від рівня значущості та кількості спостережень.

Якщо значення t -статистики перевищує критичне значення, то можна стверджувати, що коефіцієнт регресії є статистично значущим. Це означає, що відповідна незалежна змінна суттєво впливає на залежну змінну.

F-статистика використовується як метод перевірки регресійної моделі для оцінювання загальної значущості моделі загалом. Вона дозволяє визначити, чи є модель статистично значущою і чи пояснює вона значну частину варіації залежної змінної.

Значення F-статистики порівнюється з критичним значенням F розподілу (розподілу Фішера) для заданого рівня значущості. Якщо значення F-статистики перевищує критичне значення, то можна відхилити нульову гіпотезу і стверджувати, що модель є статистично значущою.

F-статистика також використовується для порівняння двох регресійних моделей. За допомогою F-статистики можна оцінити, чи є одна модель кращою за іншу, порівнюючи значення F-статистики двох моделей.

2.2 Методи непараметричної регресії для прогнозування заробітної плати

Машинне навчання – це деяка складова штучного інтелекту в галузі інформатики, яка розробляє і застосовує статистичні прийоми для надання комп'ютерам здатності навчатися: виявляти закономірності та робити висновки.

Алгоритм випадкового лісу (Random Forest) є одним з найпопулярніших методів машинного навчання для класифікації та регресії. Він комбінує принципи "ансамблювання" і "випадковості", що дає змогу покращити якість прогнозування та зменшити перенавчання моделі [23]. На відміну від методів класичної регресії метод випадкового лісу розв'язує задачу непараметричної регресії.

Алгоритм випадкового лісу будує багато дерев рішень. Кожне дерево будують на вибірці даних, яка формується випадковим вибором з повторенням з основного набору даних. Крім того, при побудові кожного дерева випадково вибираються підмножини змінних, які використовуються для розбиття вузлів дерева (рис 2.1).

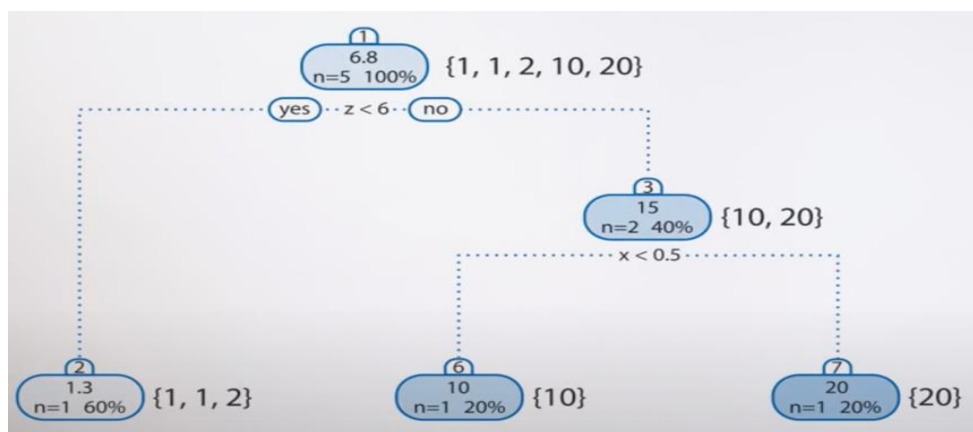


Рис. 2.1 – Приклад алгоритму Random Forest [24]

Основні кроки алгоритму випадкового лісу:

1. Формування випадкових підвибірок даних: З основного набору даних випадковим чином вибираються підмножини з повторенням. Ці підмножини будуть використовуватися для побудови окремих дерев рішень.
2. Побудова дерев рішень: На кожній підвибірці даних будується окреме дерево рішень. Кожне дерево розбиває дані на основі певних критеріїв, таких як індекс Джині для класифікації або середньоквадратична помилка для регресії.
3. Прогнозування за допомогою ансамблю: Коли всі дерева рішень побудовані, вони об'єднуються для здійснення прогнозів. У класифікації використовується голосування більшості (majority voting), де кожне дерево вносить свій вклад у прогнозовану класифікацію. У регресії результати дерев об'єднуються шляхом усереднення або голосування.
4. Оцінювання важливості ознак: Випадковий ліс надає важливість кожній ознаці (змінній), яка використовується під час побудови дерев. Це може бути використано для вибору найбільш інформативних ознак.

Переваги випадкового лісу включають стійкість до перенавчання, високу точність прогнозування, здатність обробляти великі набори даних і автоматичну внутрішню перекресну перевірку.

Цей алгоритм відзначається високою стабільністю. Навіть при додаванні нової точки даних до набору, вплив на загальний алгоритм є обмеженим, оскільки нові дані можуть впливати лише на окреме дерево, а не на всі дерева одночасно. Алгоритм випадкового лісу добре працює, коли у датасеті є як категоріальні, так і числові ознаки.

Однак, головний недолік випадкових лісів полягає в їхній складності. Вони вимагають значно більше обчислювальних ресурсів через велику кількість об'єднаних дерев рішень. До того ж в них відсутня візуалізація представлення процесу прийняття рішень щодо остаточного результату і складність інтерпретації цих рішень. Незважаючи на ці недоліки метод отримав визнання як у статистичному співтоваристві, так і серед дослідників, які розв'язують задачі класифікації та непараметричної регресії.

2.3 Кластерний аналіз та метод k-середніх

Статистичні дані щодо заробітної плати містять значну кількість показників, які характеризують ситуацію на ринку праці. Такі показники для деяких об'єктів можуть бути схожими, для інших – відмінними, тобто структура даних може бути неоднорідною. У ситуаціях, коли потрібно виявляти таку структуру – групувати дані за різними ознаками, застосовують методи кластерного аналізу.

Цей метод аналізу даних дозволяє групувати схожі об'єкти разом в кластери. Його основна мета полягає в розділенні набору даних на групи (кластери) в такий спосіб, щоб об'єкти всередині кожного кластера були подібні між собою, а об'єкти з різних кластерів були різними [22].

Основні етапи кластерного аналізу:

1. Вибір міри схожості: Спочатку потрібно визначити міру схожості (або відстані) між об'єктами даних. Це може бути, наприклад, Евклідова відстань, Манхеттенська відстань, косинусна схожість або інші відстані, залежно від типу даних і природи проблеми.

2. Вибір алгоритму: Існує багато різних алгоритмів кластерного аналізу, які можуть бути використані. Деякі з них включають метод k-середніх, ієрархічний кластерний аналіз, методи засновані на густині, спектральний аналіз кластерів тощо. Вибір алгоритму залежить від розмірів даних, природи даних, обчислювальних обмежень та мети аналізу.

3. Визначення кількості кластерів: Перед виконанням кластерного аналізу потрібно визначити кількість кластерів, на які буде розбито дані. Це може бути попередньо визначено на підставі доменних знань або може бути визначено за допомогою статистичних критеріїв або методів, таких як "ліктевий метод" (elbow method) або "метод силуету" (silhouette method).

4. Виконання кластеризації. Після вибору алгоритму і визначення кількості кластерів проводиться сам процес кластеризації. Алгоритм виконує ітерації, розташовуючи об'єкти в кластерах залежно від їх схожості, використовуючи обрану міру схожості.

5. Оцінка та інтерпретація результатів. Після завершення кластерного аналізу потрібно оцінити якість отриманих кластерів та їх інтерпретувати. Це може включати візуалізацію результатів, оцінку внутрішньокластерної і міжкластерної схожості, а також порівняння з іншими варіантами кластеризації.

Метод k-середніх є одним з найпоширеніших алгоритмів кластерного аналізу. Існує декілька алгоритмів для цього методу, однак стандартний алгоритм визначає загальну варіацію в межах кластера як суму квадратичних евклідових відстаней між елементами і відповідним центроїдом:

$$W(C_k) = \sum_{x_i \in C_k} (x_i - \mu_k)^2 \quad (2.3)$$

x_i —точка даних, що належить до кластера C_k

μ_k —середнє значення точок, що присвоєні кластеру C_k .

Кожне спостереження x_i присвоюється заданому кластеру в такий спосіб, щоб сума квадратів відстані спостереження до їх призначених кластерних центрів μ_k мінімізувалася.

Алгоритм виглядає так:

1. Визначити k — кількість кластерів, які будуть створені.
2. Вибрати випадкові k об'єктів з набору даних у ролі початкових центрів кластерів.
3. Призначити кожне спостереження найближчому центроїду на основі евклідової відстані між об'єктом і центроїдом.
4. Для кожного з k кластерів перерахувати центроїд кластера шляхом обчислення нового середнього значення усіх точок даних у кластері.
5. Ітеративно мінімізувати загальну суму в межах площі. Повторити кроки 3 і 4, поки центроїди не зміняться або не буде досягнуто максимальної кількості ітерацій (R за замовчуванням використовує 10 як максимальну кількість ітерацій).

Загальна сума, або загальна варіація в межах кластера визначається так:

$$\sum_{k=1}^k W(C_k) = \sum_{k=1}^k \sum_{x_i \in C_k} (x_i - \mu_k)^2 \quad (2.4)$$

k - кількість кластерів, які будуть створені

x_i - кожне окреме спостереження у наборі даних

μ_k - центроїд k -го кластера

2.4 Огляд технологій та середовищ для реалізації задач аналізу

Для реалізації розглянутих методів математичного моделювання створено різні програмні середовища, одним з яких є RStudio.

R - це мова програмування та середовище для статистичного аналізу та візуалізації даних.

R використовується в багатьох галузях, включаючи науку про дані, статистику, фінанси, біоінформатику, соціальні науки та машинне навчання. Мова набула популярності серед фахівців з аналізу даних завдяки великій кількості пакетів і функцій, доступних для обробки даних і візуалізації результатів.

Однією з унікальних особливостей R є її легкий та зрозумілий синтаксис. Він забезпечує зручний спосіб маніпулювання даними та виконання статистичних розрахунків. Крім того, R має потужну графічну систему, яка дозволяє користувачам візуалізувати дані в різних форматах, від простих графіків до складних інтерактивних візуалізацій.

R також має велику спільноту користувачів, що сприяє обміну знаннями та розвитку. Багато корисних пакетів доступні безкоштовно і можуть бути встановлені для розширення можливостей мови. Це дозволяє вирішувати широкий спектр завдань, від простих обчислень до складного аналізу даних і машинного навчання.

За допомогою R можна вирішувати такі завдання:

Обробка та маніпулювання даними. R надає потужні інструменти для імпорту, експорту, очищення та перетворення даних. Вона також може виконувати операції над такими об'єктами даних, як вектори, матриці, фрейми даних і списки.

Статистичний аналіз. Для статистичного аналізу даних у R існує низка пакетів. Ці пакети можна використовувати для виконання різних статистичних тестів, побудови регресійних моделей, аналізу варіації тощо.

Візуалізація даних. R дозволяє створювати високоякісні графіки для візуалізації даних: різні графічні пакети, такі як BasicGraphics, ggplot2 та

lattice, надають безліч можливостей для створення красивих та інформативних графіків.

Машинне навчання. Машинне навчання R включає пакети для машинного навчання, а саме: caret, randomForest та rpart. Ці пакети можна використовувати для виконання класифікації, регресії, кластеризації та інших завдань машинного навчання.

Отже, мова R є потужним інструментом для статистичного аналізу, візуалізації даних та машинного навчання. Її простий синтаксис, велика спільнота користувачів і наявність різноманітних пакетів роблять її популярним вибором для багатьох дослідників і аналітиків даних у всьому світі.

RStudio - це інтегроване середовище розробки (IDE) для мови програмування R, призначене для полегшення використання R шляхом надання потужних інструментів для написання коду, виконання аналізу даних, візуалізації результатів та управління проєктами. До ключових можливостей R належать:

Кодування та редактор. RStudio має зручний редактор з підсвічуванням синтаксису, автозавершенням та автоматичним форматуванням коду. Це полегшує процес написання коду, покращує його читабельність і допомагає зменшити кількість помилок.

RStudio надає інтерактивну консоль R, яка дозволяє виконувати команди R безпосередньо. Консоль забезпечує швидкий зворотній зв'язок, дозволяє частково виконувати код і відразу бачити результати. Історія команд також зберігається, тому попередні операції можна легко повторити або змінити.

Управління проєктами. RStudio дозволяє створювати та керувати проєктами, організовуючи код, дані та результати в логічні групи. Програміст може виконувати дії, пов'язані з організацією його роботи, такі як створення нових проєктів, керування пакетами та налаштуваннями проєкту.

В RStudio інтегровано інструменти для побудови графіків та візуалізації даних. Користувачі можуть створювати різні типи графіків, задавати параметри графіків та маніпулювати візуалізацією безпосередньо з відповідного редактора.

Інтеграція з пакетами та інструментами. RStudio має вбудовану підтримку для керування пакетами R, що дозволяє встановлювати, оновлювати та завантажувати пакети безпосередньо з інтерфейсу. Він також інтегрується з іншими популярними інструментами, такими як Git для контролю версій, що полегшує роботу з командним рядком і обліковими записами.

RStudio є дуже популярним середовищем розробки R, оскільки покращує продуктивність, ефективність та якість коду, а також полегшує роботу з даними та їх аналіз в R.

РОЗДІЛ 3

ПРАКТИЧНЕ ЗАСТОСУВАННЯ МОДЕЛЬНОГО ІНСТРУМЕНТАЛЬНОГО АНАЛІЗУ ДО ЗАРОБІТНОЇ ПЛАТИ ІТ ФАХІВЦЯ

3.1 Модель прямої покрокової регресії

В роботі використано дані з опитування найбільшої ІТ-спільноти в Україні DOU. З 2010 року, кожні 6 місяців проводяться опитування серед людей, які працюють на ІТ ринку та публікуються ці дані у вільному доступі [25].

Набір даних складається з інформації щодо посади, заробітної плати, міста, в якому працює найманий працівник, спеціалізації, мови програмування, яку використовує програміст на щоденній основі, віку, наявності вищої освіти, рівня знань англійської мови, досвіду роботи за фахом та у компанії. Розглядаються ІТ-спеціалісти, які використовують мову Javascript.

В датасеті представлені такі кількісні та якісні дані (рис 3.1):

Salary – зарплата фахівця (дол.США на рік);

City – місто, в якому проживає респондент;

Gender – стать;

Education – інформацію про наявність вищої освіти;

University – університет, в якому навчався респондент;

English – рівень знання англійської мови;

Sphere – напрямок компанії, в якій працює респондент;

Type_Of_Company – тип компанії, в якій працює респондент;

Number_Of_Coworkers – кількість співробітників у компанії, в якій працює респондент;

Total_Experience – досвід роботи у галузі (у роках);

Current_Exp – досвід роботи на поточному місці (у роках);

Age – вік респондента (у роках).

Результативною змінною є заробітна плата. До факторів, які впливають на заробітну плату, можна віднести: загальний досвід роботи, досвід роботи на поточному місці, знання англійської мови, вік, тип компанії, та напрямок компанії.

```
> summary(salaries)
  City          Gender          Age          Education          University          English
Length:1244    Length:1244    Min.   :17.00    Length:1244    Length:1244    Length:1244
Class :character Class :character 1st Qu.:24.00    Class :character Class :character Class :character
Mode  :character Mode  :character Median :27.00    Mode  :character Mode  :character Mode  :character
                        Mean  :27.68
                        3rd Qu.:31.00
                        Max.   :50.00
 Sphere      Type_Of_Company  Number_Of_Coworkers  Total_Exp    Current_Exp    Salary
Length:1244  Length:1244    Length:1244          Min.   : 0.250 Min.   : 0.250 Min.   : 150
Class :character Class :character Class :character 1st Qu.: 2.000 1st Qu.: 0.500 1st Qu.: 1000
Mode  :character Mode  :character Mode  :character Median : 3.000 Median : 1.000 Median : 2100
                        Mean  : 3.887 Mean  : 1.551 Mean  : 2397
                        3rd Qu.: 5.000 3rd Qu.: 2.000 3rd Qu.: 3500
                        Max.   :10.000 Max.   :10.000 Max.   :10000
```

Рисунок 3.1 – Загальна інформація про дані

Джерело: розроблено автором на основі [27]

Змінні Total_Exp, Current_Exp, Salary та Age належать до числового типу даних. Змінні City, Gender, Sphere, Type_Of_Company, Education, University та English належать до типу даних character.

Для визначення зв'язку між змінними побудовано кореляційну матрицю (рис. 3.2):

	salary	Total_Exp	Current_Exp	Age
salary	1.0000000	0.7433044	0.2810772	0.3336932
Total_Exp	0.7433044	1.0000000	0.4963372	0.4738207
Current_Exp	0.2810772	0.4963372	1.0000000	0.2351962
Age	0.3336932	0.4738207	0.2351962	1.0000000

Рисунок 3.2 – Кореляційна матриця

Джерело: розроблено автором на основі [28]

Значення кореляції може бути від -1 до 1. Значення, близькі до 1, вказують на прямий лінійний зв'язок, тоді як значення, близькі до -1, вказують на обернений лінійний зв'язок. Значення близькі до 0 свідчать про відсутність лінійного зв'язку

між змінними. Великі значення кореляції (близькі до 1 або -1) вказують на сильний лінійний зв'язок між змінними. Менші значення (близькі до 0) свідчать про слабкий зв'язок. Аналізуючи кореляційну матрицю, можна дійти висновку про сильну кореляцію між загальним досвідом роботи та заробітною платнею спеціаліста. Тобто, чим більший досвід роботи у фахівця, тим більшою є його заробітна плата. Також існує слабкий лінійний зв'язок між досвідом роботи у поточній компанії та заробітною платнею фахівця та між віком і досвідом роботи. Помірний зв'язок існує між загальним досвідом роботи із досвідом роботи на поточній посаді та між загальним досвідом роботи і з віком фахівця. Зі збільшенням досвіду роботи, збільшується і вік фахівця та його стаж на поточному місці роботи.

Після отримання таких результатів, доцільним буде побудова прямої покрокової регресії, через наявність кореляції між змінними salary та Total_Exp, Current_Exp, Age.

Пряма покрокова регресія реалізована за допомогою функції `step()` та пошукового алгоритму «forward stepwise». Алгоритм додає змінні по одній на кожній ітерації, починаючи з найбільш значущої змінної, на кожному кроці оцінюючи вплив нової змінної на модель і включаючи її, якщо це покращує якість моделі. Аргумент `score` визначає множину змінних, з якої можуть бути включені змінні, і `trace = 0` припиняє роботу алгоритму. За допомогою функції `summary()` визначаються основні статистики і коефіцієнти моделі.

Результати моделі `fwd.model` показують значущий вплив змінних Total_Exp і Current_Exp на заробітну плату (`salaries$Salary`). Коефіцієнти при змінних моделі показують, що з кожною одиницею збільшення `salaries$Total_Exp`, заробітна плата збільшується в середньому на 463.79 одиниць. Коефіцієнти статистично значущі, оскільки мають низькі значення `p-value` ($< 2.2e-16$). Модель не включає змінну `salaries$Age`, оскільки її вплив не є статистично значущим. Коефіцієнт детермінації Multiple R-squared дорівнює 0.5627, що означає, що модель пояснює близько 56.27% варіації заробітної плати. F-статистика має значення 798.6 і показує, що модель загалом є статистично значущою (рис.3.3).

```
> summary(fwd.model)

Call:
lm(formula = salaries$Salary ~ salaries$Total_Exp + salaries$Current_Exp)

Residuals:
    Min       1Q   Median       3Q      Max
-4980.7  -583.5  -136.9   530.0  7253.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)      771.46      52.15  14.794 < 2e-16 ***
salaries$Total_Exp   463.79      12.52  37.053 < 2e-16 ***
salaries$Current_Exp -114.50      21.24  -5.391 8.37e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1033 on 1241 degrees of freedom
Multiple R-squared:  0.5627,    Adjusted R-squared:  0.562
F-statistic: 798.6 on 2 and 1241 DF,  p-value: < 2.2e-16
```

Рисунок 3.3 – Результати роботи функції summary() для моделі fwd.model

Джерело: розроблено автором на основі [27]

Отже, отримана модель включає статистично значущі змінні та добре пояснює частину варіації в заробітній платі. Проте для повної оцінки моделі потрібно провести ще багато аналізів.

3.2 Вплив якісних показників на заробітну плату

Кваліфікація фахівця будь-якої галузі формується насамперед в певному освітньому середовищі. У суспільстві існує дилема про потребу наявності вищої освіти для ІТ спеціаліста. Низка людей дотримується думки, що для роботи в ІТ обов'язковою є вища освіта, особливо якщо кандидат планує працювати розробником. Рекрутери приймають це до уваги під час найму, особливо в аутсорс та аутстаф компаніях, які працюють на експорт. Замовник часто потребує від учасників команд наявність вищої освіти, тому компанії такого типу наймають кандидатів, спираючись на цей фактор. Та чи впливає освіта на заробітну плату фахівця?

Для відповіді на це запитання, була побудована модель m_education, в якій salary є залежною змінною, а Total_Exp, Current_Exp та Education – незалежними.

Кореляція між змінними Education та salary склала 0.0195681, а результати побудови цієї моделі подано на рис. 3.4:

```
Call:
lm(formula = Salary ~ Total_Exp + Current_Exp + Education, data = salaries)

Residuals:
    Min       1Q   Median       3Q      Max
-4991.4  -621.9  -158.9   565.6  7246.2

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    825.05      63.94  12.904 < 2e-16 ***
Total_Exp       459.72      14.61  31.467 < 2e-16 ***
Current_Exp    -123.28      23.59  -5.226 2.11e-07 ***
Education       112.70      98.73   1.142  0.254
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1076 on 1002 degrees of freedom
Multiple R-squared:  0.5298,    Adjusted R-squared:  0.5284
F-statistic: 376.4 on 3 and 1002 DF,  p-value: < 2.2e-16
```

Рисунок 3.4 – Результати роботи функції summary() для моделі m_education

Джерело: розроблено автором на основі [27]

Значення t-статистики (t-value) для Education дорівнює 1.142. Значення цієї статистики вказує на відносно низьку статистичну значущість змінної Education. Значення p-рівня значущості ($\Pr(>|t|)$) для Education становить 0.254, що більше відповідного значення 0.05. Це означає, що немає статистично значущих доказів впливу освіти на заробітну плату в цій моделі.

Отже, на основі результатів можна зробити висновок, що освіта не є статистично значущою незалежною змінною для моделі заробітної плати. Це означає, що враховуючи загальний стаж роботи та досвід роботи в поточній компанії, освіта не впливає значимо на рівень заробітної плати. Проте, без вищої освіти кандидату буде набагато важче працевлаштуватись, тому освіта є також важливою перевагою, яку повинен кандидат для успішного працевлаштування.

Англійська мова – мова бізнесу. Абсолютна більшість міжнародних комунікацій проходить англійською мовою і це безпосередньо стосується ІТ-бізнесу. Більшість компаній в Україні працюють на європейський/американський

ринок і роботодавець розраховує на те, щоб розробники також підтримували комунікацію із замовниками. Також спілкування між співробітниками в міжнародних ІТ-компаніях проходить англійською мовою. Всі ці перелічені причини говорять про важливість знання англійської мови для розробників.

Для того, аби оцінити вплив знання англійської мови на зарплату кандидата, була побудована модель `m_english` (рис. 3.5):

```
> m_english <- lm(Salary~Total_Exp+Current_Exp+English, data=salaries)
> summary(m_english)
```

Call:
lm(formula = Salary ~ Total_Exp + Current_Exp + English, data = salaries)

Residuals:

Min	1Q	Median	3Q	Max
-4664.8	-578.1	-114.8	547.6	6985.2

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	25.83	96.50	0.268	0.789
Total_Exp	424.51	14.23	29.828	< 2e-16 ***
Current_Exp	-123.58	22.36	-5.527	4.16e-08 ***
English	412.38	38.50	10.710	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1020 on 1002 degrees of freedom
Multiple R-squared: 0.5776, Adjusted R-squared: 0.5763
F-statistic: 456.7 on 3 and 1002 DF, p-value: < 2.2e-16

Рисунок 3.5 – Результати роботи функції `summary()` для моделі `m_english`

Джерело: розроблено автором на основі [27]

Коефіцієнт оцінки (Estimate) для English дорівнює 412.38. Це означає, що зі збільшенням володіння англійською мовою на рівень, очікуване значення заробітної плати збільшується на 412.38 одиниць. Значення t-статистики (t-value) для English дорівнює 10.710. Значення цієї статистики є статистично значущим, оскільки перевищує критичне значення (зазвичай при рівні значущості 0.05). Значення p-рівня значимості ($\text{Pr}(>|t|)$) для English дуже мале ($< 2.2e-16$), що вказує на статистичну значущість змінної English.

Отже, на основі результатів можна зробити висновок, що рівень володіння англійською мовою (English) має значний вплив на рівень заробітної плати.

При пошуку роботи, кандидати в свою чергу звертають увагу на розмір ІТ компанії. Маленькі компанії зазвичай пропонують більш гнучкий графік роботи, робота найманого працівника буде охоплювати різні аспекти, що дає можливість навчитися різноманітним навичкам. У великих компаніях більший обсяг проєктів та більша команда розробників, що відкриває можливості до нетворкінгу, новим знайомствам та обміну досвідом. Та чи впливає розмір компанії на заробітну плату кандидата?

Для дослідження цього питання було побудовано модель `m_coworkers`, в якій `salary` є залежною змінною, а `Total_Exp`, `Current_Exp` та `Number_Of_Coworkers` – незалежними.

У даному разі, оцінка коефіцієнта для `Number_Of_Coworkers` дорівнює 69.93, а стандартна помилка оцінки - 25.29. Значення t-статистики для цього коефіцієнта дорівнює 2.765, а p-value дорівнює 0.00581.

Оскільки p-value менше ніж 0.05 (за прийнятим рівнем значущості), то можна стверджувати, що змінна `Number_Of_Coworkers` є статистично значущою для моделі. Це означає, що кількість співробітників (`Number_Of_Coworkers`) має статистично значущий вплив на заробітну плату (`Salary`), при урахуванні інших змінних у моделі (`Total_Exp` і `Current_Exp`). І при збільшенні кількості працівників у компанії, збільшується і заробітна плата працівника (рис. 3.6)

```
> m_coworkers <- lm(Salary~Total_Exp+Current_Exp+Number_Of_Coworkers, data=salaries)
> summary(m_coworkers)
```

Call:
lm(formula = Salary ~ Total_Exp + Current_Exp + Number_Of_Coworkers,
data = salaries)

Residuals:

Min	1Q	Median	3Q	Max
-4900.3	-634.0	-143.0	552.4	7243.9

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	648.13	92.76	6.987	5.12e-12 ***
Total_Exp	454.35	14.69	30.927	< 2e-16 ***
Current_Exp	-124.53	23.52	-5.295	1.46e-07 ***
Number_Of_Coworkers	69.93	25.29	2.765	0.00581 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1073 on 1002 degrees of freedom
Multiple R-squared: 0.5328, Adjusted R-squared: 0.5314
F-statistic: 380.9 on 3 and 1002 DF, p-value: < 2.2e-16

Рисунок 3.6 – Результати роботи функції `summary()` для моделі `m_coworkers`

Джерело: розроблено автором на основі [27]

Для подальшого дослідження було обрано модель `m_english`, тому що вона показала найбільший R^2 , який дорівнює 0,58. Це означає, що 58% варіації залежної змінної може бути пояснено використаними незалежними змінними у моделі. Для початку було обчислено залишки моделі та виведено статистичний підсумок цих залишків за допомогою функції `summary()`. Залишки є різницею між спостереженими та прогнозованими значеннями, отриманими з моделі.

Аналізуючи результати, які представлені на рис. 3.7 можна дійти висновку про те, що у даній моделі помилки розподілені навколо нуля, що вказує на низький рівень систематичних відхилень. Наступним кроком є побудова діаграми залишків.

```
> summary (error)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-4664.8	-578.1	-114.8	0.0	547.6	6985.2

Рисунок 3.7 – результати роботи функції `summary()` за залишками

Джерело: розроблено автором на основі [27]

На рис. 3.8 форма діаграми залишків моделі подібна до нормального розподілу, більшість залишків зосереджена навколо нуля. Гістограма має трохи асиметричну форму, але це не критично. Це говорить про те, що більшість залишків є від'ємними. Є деякі значення залишків, які виходять за межі нормального розподілу. Це може свідчити про наявність викидів або потенційних аномалій у даних. Загалом, гістограма залишків свідчить про те, що модель добре підходить для більшості спостережень і здатна відтворювати дані. Однак, наявність деяких значень за межами нормального розподілу може вказувати на те, що модель не враховує деякі важливі фактори або що в них присутні випадкові аномалії.

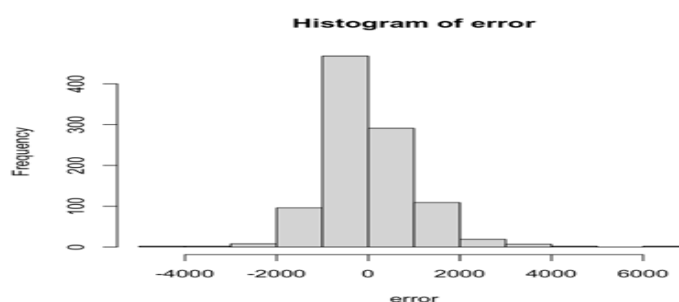


Рисунок 3.8 – гістограма похибок

Джерело: розроблено автором на основі [27]

Функція outlierTest (m_english) вказала на спостереження під номерами 1, 2, 8, 984 та 841, які є викидами.

Графік на рис. 3.9 відображає відношення залишкових значень до прогнозованих значень. Він допомагає оцінити наявність нелінійності, гетероскедастичності та викидів у моделі. Він показує, що більшість залишків розташовані випадково навколо горизонтальної лінії, що свідчить про адекватність моделі.

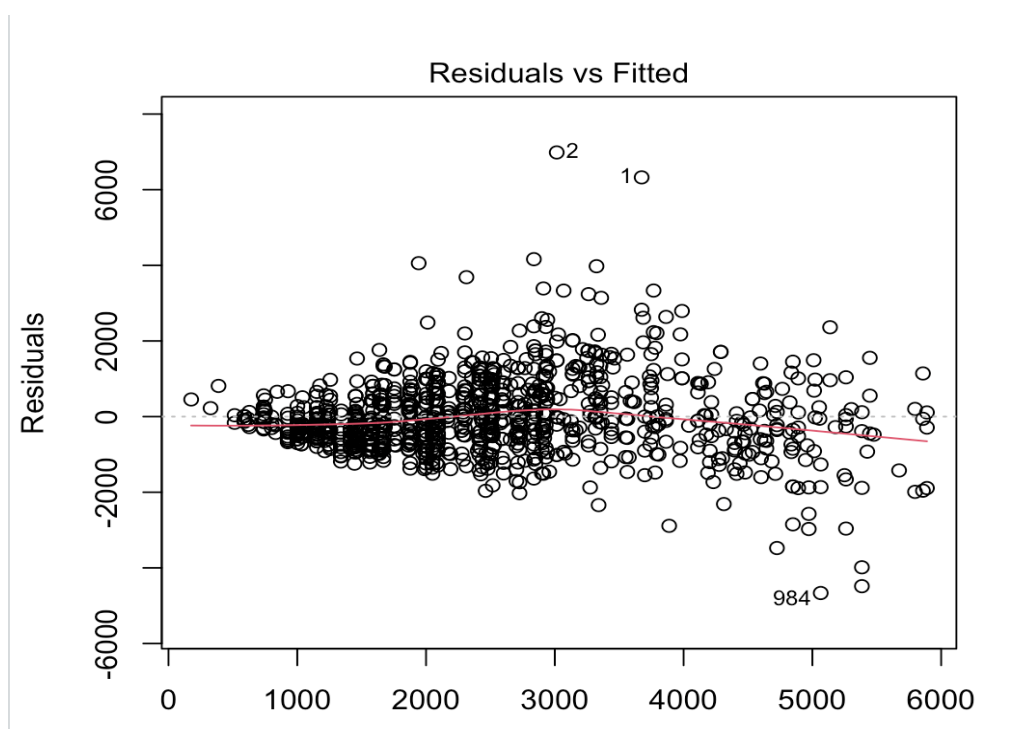


Рисунок 3.9 – графік Residuals vs Fitted

Джерело: розроблено автором на основі [27]

Normal Q-Q. Цей графік використовується для перевірки нормальності розподілу залишкових значень. В центральній частині графіка на рис. 3.10 спостерігається більша концентрація точок ближче до пунктирної лінії, яка є лінією нормального розподілу, а на обох кінцях графіка є деякі відхилення від лінійної форми. На підставі цього графіка можна припустити, що розподіл залишків має деяку нормальну або близьку до нормальної форму, з деякими відхиленнями на кінцях. Однак, для більш точних висновків необхідно провести додатковий аналіз та врахувати інші статистичні метрики.

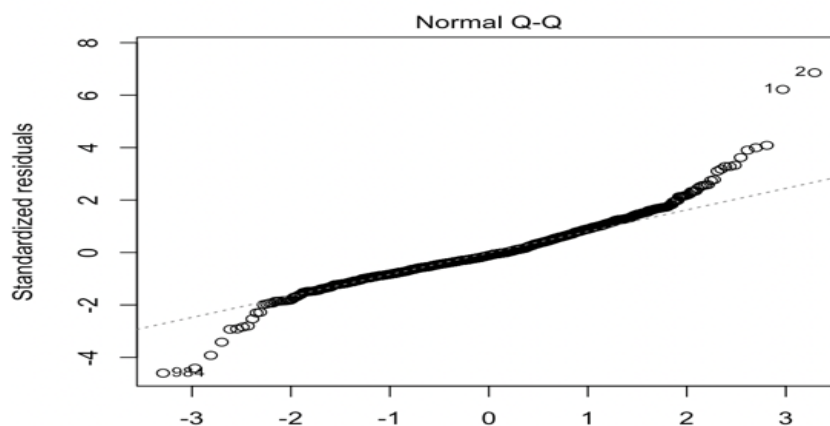


Рисунок 3.10 – графік Normal Q-Q

Джерело: розроблено автором на основі [27]

На рис. 3.11 зображені стандартизовані залишки (залишки, поділені на їхнє стандартне відхилення) по відношенню до прогнозованих значень. Він допомагає виявити викиди або незвичайні спостереження, які можуть впливати на модель. Абсолютна більшість стандартизованих залишків знаходяться в межах ± 2 , це може свідчити про адекватність моделі.

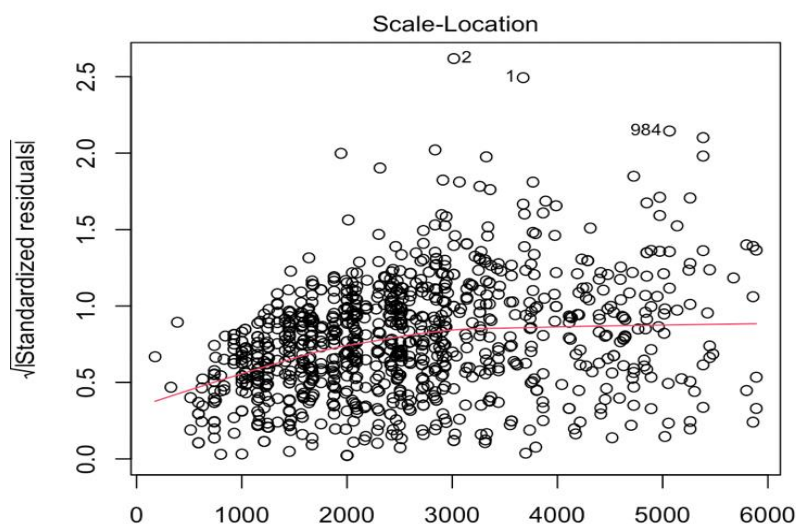


Рисунок 3.11 – графік Scale-Location

Джерело: розроблено автором на основі [27]

На рис. 3.12 зображені залишки моделі відносно їхніх прогнозованих значень. Він допомагає виявити нелінійність, гетероскедастичність або інші систематичні закономірності у залишках. Залишки розподілені випадково і рівномірно навколо нуля, це може свідчити про відповідність моделі даним.

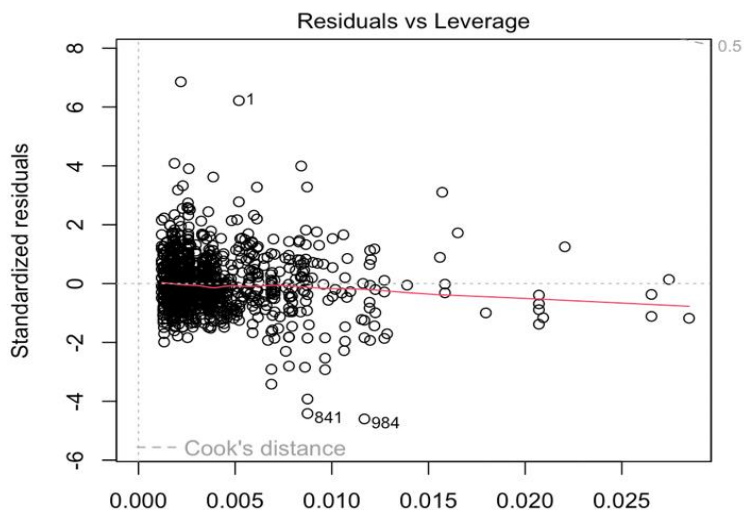


Рисунок 3.12 – графік Residuals vs Leverage

Джерело: розроблено автором на основі [27]

Загалом, на підставі побудованих графіків можна стверджувати, що модель відповідає основним лінійним припущенням та має задовільні статистичні характеристики. Однак, на всіх графіках були видні ті ж самі викиди, які були знайдені за допомогою `outlierTest(m_english)`. Тому вирішено видалити ці викиди з датасету та створити нову модель без них. Нова модель отримала назву `new_m_english`. Та мала $R^2 = 0,62$, що на 4% більша, ніж модель з викидами.

Проте, навіть після видалення викидів з моделі, вона не описує достатню кількість варіації. Саме тому була побудована модель з використанням логарифмів `model_log <- lm(log(Salary)~log(Total_Exp), data=new_salaries)` та проведена оцінка моделі за допомогою функції `summary()`.

Результати оцінювання моделі на рис. 3.13 свідчать про те, що модель `model_log` має коефіцієнт детермінації R-квадрат, який дорівнює 0.6858. Це означає, що приблизно 68.58% варіації логарифму заробітної плати можна пояснити за допомогою логарифму загального досвіду в моделі. Коefіцієнти мають низькі p-value ($< 2.2e-16$). Це свідчить про статистичну значимість коефіцієнтів у моделі. Значення F-статистики для моделі `model_log` дорівнює 2700 з 1 і 1237 ступенями свободи. Високе значення F-статистики та низьке p-value підтверджують, що модель в цілому є статистично значущою.

```
> summary(model_log)
```

Call:

```
lm(formula = log(Salary) ~ log(Total_Exp), data = new_salaries)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.43048	-0.25642	0.03868	0.29550	1.14080

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.78038	0.01882	360.24	<2e-16 ***
log(Total_Exp)	0.71292	0.01372	51.96	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4279 on 1237 degrees of freedom

Multiple R-squared: 0.6858, Adjusted R-squared: 0.6855

F-statistic: 2700 on 1 and 1237 DF, p-value: < 2.2e-16

Рисунок 3.13 – Результати роботи функції `summary()` для моделі `model_log`

Джерело: розроблено автором на основі [27]

Для визначення впливу мовного фактора на рівень заробітної плати до побудованої моделі `model_log` додано факторну змінну рівня знань англійської мови. Модель отримала назву `model_log_eng`, яка у програмному коді має вигляд: `model_log_eng <- lm ( log ( Salary ) ~ log ( Total_Exp ) + English, data = new_salaries)`.

Результати оцінювання моделі, подані на рис. 3.14, свідчать про те, що модель описує 72% варіації. Всі коефіцієнти моделі є статистично значущими, а значення F-статистики дорівнює 1562 та має надзвичайно мале значення p-value (< 2.2e-16), що свідчить про статистичну значущість моделі загалом.

```
> summary(model_log_eng)
```

Call:

```
lm(formula = log(Salary) ~ log(Total_Exp) + English, data = new_salaries)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-2.3398	-0.2460	0.0397	0.2899	1.1466

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	6.45637	0.03325	194.18	<2e-16 ***
log(Total_Exp)	0.67607	0.01342	50.37	<2e-16 ***
English	0.15909	0.01376	11.56	<2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4066 on 1236 degrees of freedom

Multiple R-squared: 0.7164, Adjusted R-squared: 0.716

F-statistic: 1562 on 2 and 1236 DF, p-value: < 2.2e-16

Рисунок 3.14 – Результати роботи функції `summary()` для моделі `model_log`

Джерело: розроблено автором на основі [27]

Дискримінація жінок - це гостре питання, яке піднімається в інфополі дуже часто. Майже щоденно з'являються пости про нерівність оплати праці між чоловіками та жінками, тобто при однакових навичках, досвіді та виконання тієї ж кількості роботи, жінки можуть отримувати меншу заробітну плату, порівняно із чоловіками.

Для дослідження цього питання спочатку було перетворено значення стовпця Gender датасету salaries в факторні змінні. Значення 0 відповідає чоловічій статі, а 1 – жіночій. Кореляція між зарплатою та статтю склала -0,12. Негативне значення кореляції вказує на слабку обернену залежність між статтю та заробітною платнею. Попередньо можна сказати про те, що дискримінації майже не існує.

Для детальнішого дослідження цього питання була побудована модель m_descrimination, в якій salary є залежною змінною, а Total_Exp, Current_Exp та Gender – незалежними (рис. 3.15)

```
> m_descrimination <- lm(Salary~Total_Exp+Current_Exp+Gender, data=salaries)
> summary(m_descrimination)
```

Call:
lm(formula = Salary ~ Total_Exp + Current_Exp + Gender, data = salaries)

Residuals:

Min	1Q	Median	3Q	Max
-4980.7	-594.7	-130.7	536.7	7227.7

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	805.69	53.94	14.936	< 2e-16 ***
Total_Exp	460.29	12.58	36.599	< 2e-16 ***
Current_Exp	-111.61	21.23	-5.257	1.72e-07 ***
Gender	-226.25	93.72	-2.414	0.0159 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1031 on 1240 degrees of freedom
Multiple R-squared: 0.5648, Adjusted R-squared: 0.5637
F-statistic: 536.4 on 3 and 1240 DF, p-value: < 2.2e-16

Рисунок 3.15 – Результати роботи функції summary() для моделі
m_descrimination

Джерело: розроблено автором на основі [27]

Коефіцієнт оцінки (Estimate) для Gender дорівнює -226.25. Це означає, що середнє значення заробітної плати для жінок в цьому наборі даних менше на 226.25 одиниць порівняно з чоловіками. Значення t-статистики (t-value) для Gender

дорівнює -2.414. Це вказує на те, що вплив гендеру на заробітну плату є статистично значущим. Отже, на основі результатів можна зробити висновок, що гендер є статистично значущою незалежною змінною для моделі заробітної плати. Отримані результати вказують на наявність різниці у заробітній платі між чоловіками та жінками, причому середнє значення заробітної плати для жінок є меншим. Тобто за спостереженими даними виявлена дискримінація за статтю.

3.3 Порівняння результатів прогнозування моделей

Модель випадкового лісу, яка реалізовує непараметричну регресію, не має жодної візуалізації, а результати моделювання порівнюються з деякою частиною вибірових спостережень. Для роботи алгоритму випадкового лісу початкова вибірка даних розбивається на дві нерівні частини: 75% спостережень утворюють навчальну вибірку і 25% - тестову, на якій перевіряється побудована модель.

Random Forest є композицією безлічі вирішальних дерев, що дозволяє знизити проблему перенавчання і підвищити точність порівнянно з одним деревом.

Основна функція для побудови такого дерева називається `randomForest()`, та має вигляд

```
model_rf <- randomForest( Salary ~ Total_Exp + Current_Exp + English, data =
new_salaries)
```

Результати застосування цієї функції, представлені на рис. 3.16, вказують на те, що відсоток пояснених значень становить 65%.

```
Call:
randomForest(formula = Salary ~ Total_Exp + Current_Exp + English,      data = new_salaries)
      Type of random forest: regression
      Number of trees: 500
No. of variables tried at each split: 1

      Mean of squared residuals: 820383.3
      % Var explained: 64.82
```

Рисунок 3.16 – Результати роботи модуля випадкового лісу

Джерело: розроблено автором на основі [27]

Для порівняння цього методу прогнозування з іншими моделями, було обрано ще 3 моделі, які за попередніми тестами показували найкращі результати. Цими моделями стали `model_lm`, `model_rf`, `model_log` та `model_log_eng`. Навчальна та тестова вибірки для усіх моделей є однаковими та було взяті на основі датасету `new_salaries`.

У табл. 3.1 побудовані моделі порівнюються за показниками RMSE та MAPE, які визначають точність прогнозування цих моделей. Чим менші ці показники, тим точнішою є модель. Всі моделі мають показники RMSE близькі до 450 MAPE та близькі до 9,5. Значення MAPE менші 10% свідчать про високу точність прогнозування, а ті, що не перевищують 20% - про досить високу точність.

Таблиця 3.1 – Порівняння результатів прогнозування побудованих моделей.

		RMSE	MAPE
Random Forest	<code>model_rf</code>	433.97	10.33
<code>lm((log(Salary)~log(Total Exp))</code>	<code>model_log</code>	499.29	9.25
<code>lm((log(Salary)~log(Total Exp)+English)</code>	<code>model_log_eng</code>	467.05	8.54
<code>lm(Salary ~ Total Exp+Current Exp+English)</code>	<code>model_lm</code>	472.28	10.84

Джерело: розроблено автором на основі [30]

3.4 Кластерний аналіз

Для проведення кластерного аналізу був створений новий датасет, який складається лише з однієї змінної - заробітна плата (`salary`). Далі, за допомогою статистичного методу k-середніх було визначено оптимальну кількість кластерів, на які слід розбити спостереження. Для цього була використана функція $W(k)$, яка вимірює внутрішньо-груповий розкид в залежності від кількості кластерів.

Побудований графік (рис. 3.17) демонструє залежність між кількістю кластерів (k) і значенням функції $W(k)$. Цей графік нагадує криву, і при певному значенні k спостерігається згин. У даному випадку, коліно згинається при $k=3$.

Враховуючи це, можна припустити, що розбиття спостережень на 3 кластери надає достатньо внутрішньо однорідних груп, які відрізняються між собою.

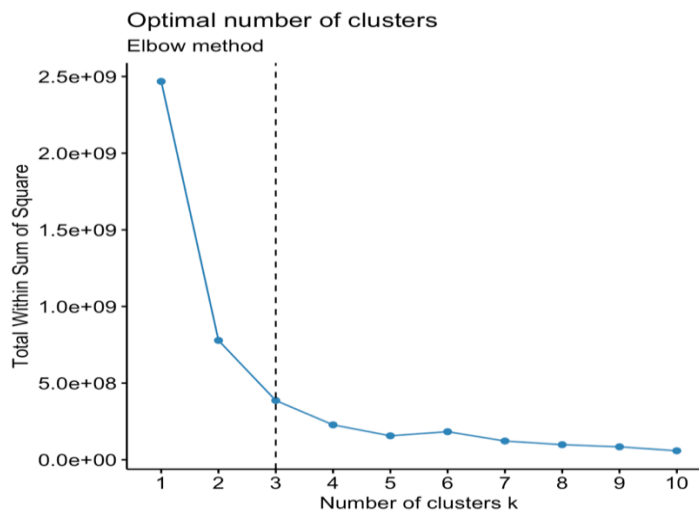


Рисунок 3.17 – Графік k-середніх

Джерело: розроблено автором на основі [29]

Побудована кластерна дендограма показує, що обґрунтованим є розбиття спостережень на 3 кластери.

Кластери на рис. 3.18 можна інтерпретувати як розбиття спеціалістів на три групи такі як Junior, Middle та Senior.

Cluster Dendrogram

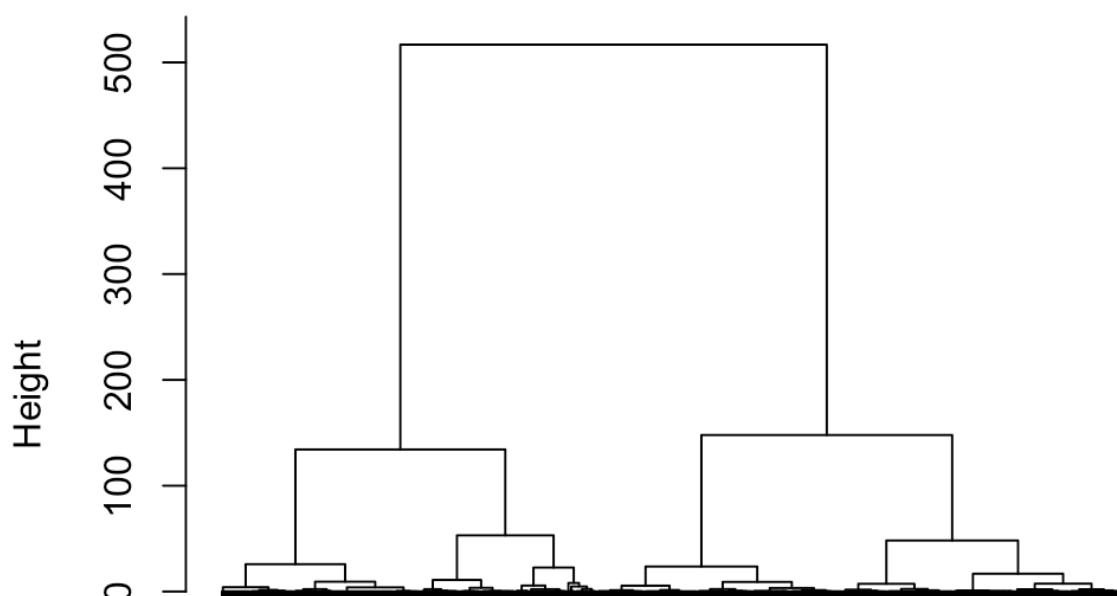


Рисунок 3.18 – Кластерна дендограма

Джерело: розроблено автором на основі [29]

Побудувавши boxplot діаграми, які представлені на рис. 3.19, можна дійти висновку про те, що розподіли заробітної плати в різних кластерах зовсім не схожі один на одного: відрізняються не тільки медіана, але і розкид значень. Найменший розкид значень має 3-тя група, яку можна інтерпретувати як Junior, а найбільший розкид – 1-ша група – Senior. Єдина група, яка має викиди – Senior і це пояснюється тим, що існують одиничні фахівці, які мають заробітну плату, значно вищу ніж у абсолютної більшості інших розробників.

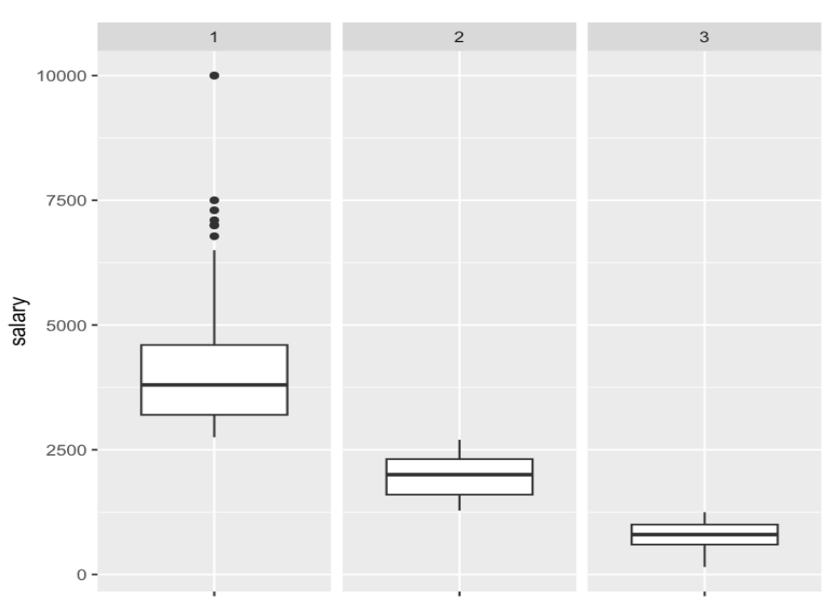


Рисунок 3.19 – boxplot діаграми

Джерело: розроблено автором на основі [29]

На графіку, який представлений на рис. 3.20, можна побачити цікаві особливості розподілу. В першій групі, основні значення заробітної плати сконцентровані на відмітці 2700-3700\$. Помітне зниження кількості спостережень, одразу після 5000\$. У другій групі майже немає розподілу, значення заробітної плати зосереджені в проміжку між 1250\$-2500\$. Третя група має розмах від 300-400\$ до 1250\$. Абсолютна більшість значень знаходиться в проміжку 800\$-1250\$.

Ці особливості розподілу заробітної плати в різних групах можуть свідчити про наявність різних категорій або сегментів заробітної плати в досліджуваній вибірці. Глибокий аналіз цих груп та дослідження факторів, що впливають на

заробітну плату в кожній групі, можуть допомогти краще зрозуміти варіацію заробітних плат в ІТ.

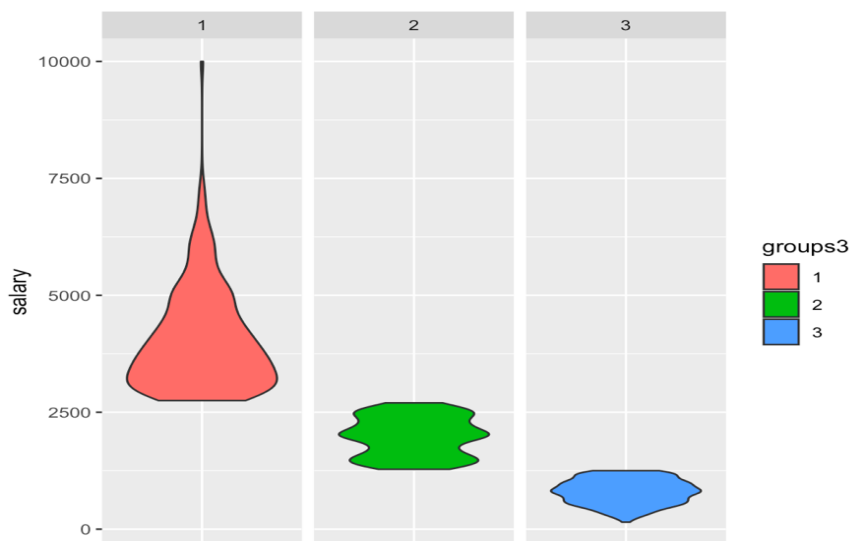


Рисунок 3.20 – violin plot діаграма

Джерело: розроблено автором на основі [29]

На графіках, які можна побачити на рисунках 3.21-3.23, відображена залежність між досвідом роботи та заробітною платою ІТ-спеціалістів. Кожен графік відповідає одному з трьох кластерів, а кольори на графіку вказують на рівень володіння англійською мовою: червоний відповідає рівню Elementary, а синій - Advanced.

За допомогою візуалізації можна помітити, що зі зростанням досвіду та заробітної плати ІТ-спеціалістів зростає й рівень володіння англійською мовою. Більшість спеціалістів з кластера, який інтерпретується як Junior, мають рівень Intermediate знання англійської мови. У Middle спеціалістів переважає рівень Intermediate та Upper-Intermediate, а Senior розробники володіють рівнями Upper-Intermediate та Advanced.

Це підкреслює важливість знання англійської мови для ІТ-спеціалістів і свідчить про те, що без володіння цією мовою стати фахівцем в цій галузі майже неможливо. Знання англійської мови допомагає спілкуватися з колегами з інших країн, отримувати нові знання та просуватися у кар'єрному розвитку в ІТ-сфері.

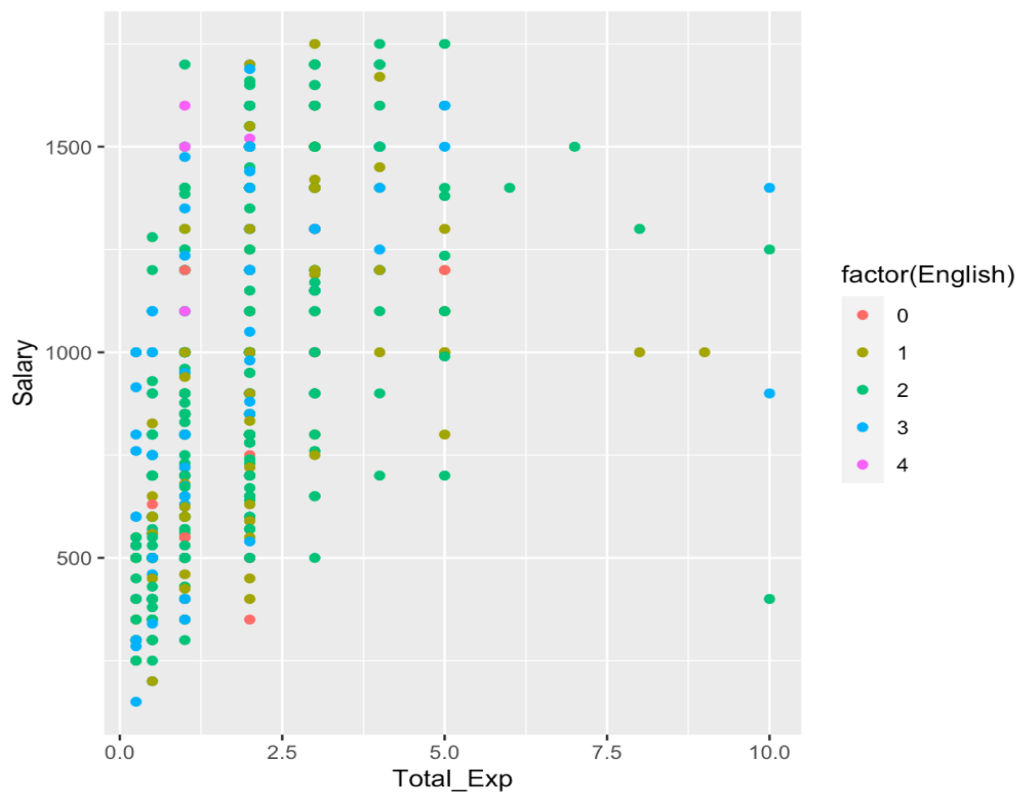


Рисунок 3.21 – Діаграма залежності між заробітною платою та загальним досвідом роботи для 1-го кластеру (Junior) з урахуванням рівня англійської мови
Джерело: розроблено автором на основі [29]

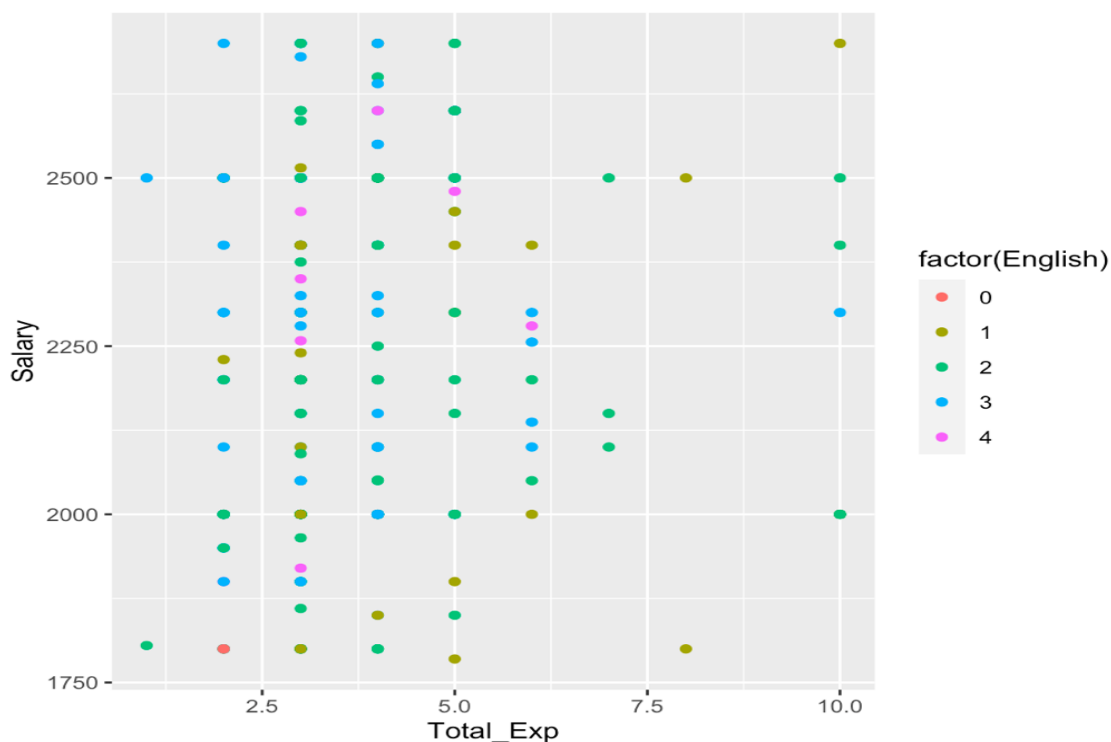


Рисунок 3.22 – Діаграма залежності між заробітною платою та загальним досвідом роботи для 2-го (Middle) кластеру з урахуванням рівня англійської мови
Джерело: розроблено автором на основі 29]

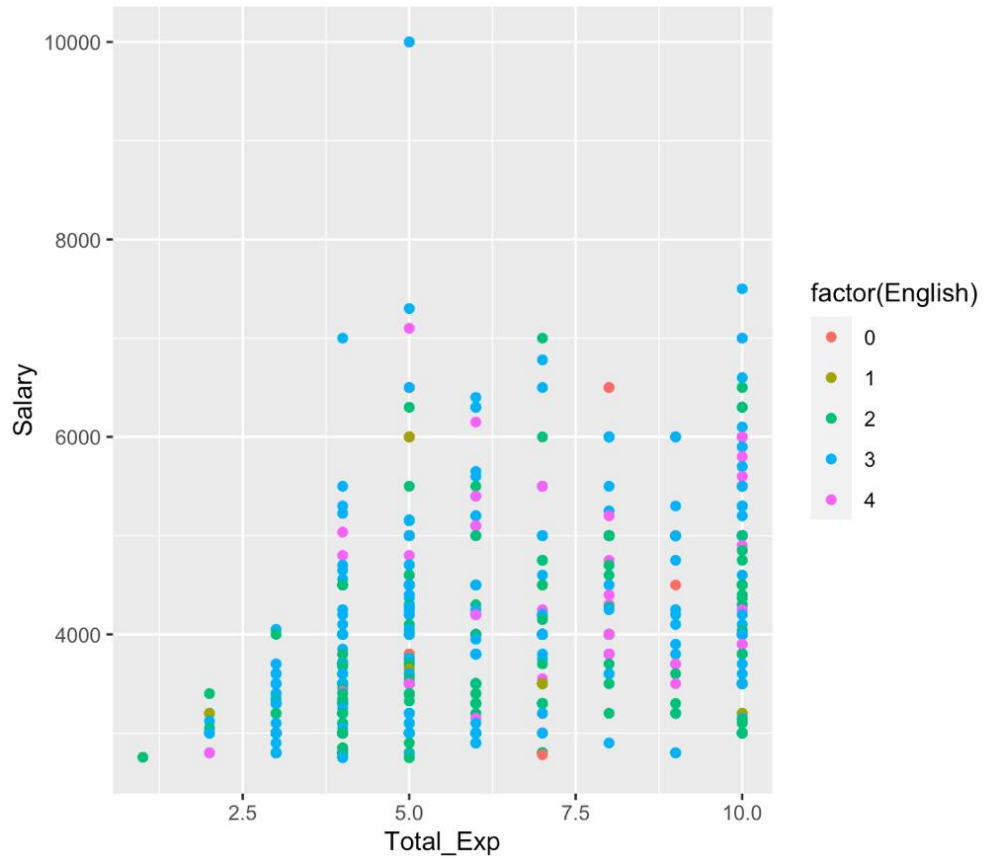


Рисунок 3.23 – Діаграма залежності між заробітною платою та загальним досвідом роботи для 3-го кластеру (Senior) з урахуванням рівня англійської мови
Джерело: розроблено автором на основі [29]

ВИСНОВКИ

ІТ-галузь в Україні за останні роки стала провідною з експорту послуг, а українські ІТ-спеціалісти здобули репутацію висококваліфікованих професіоналів. Через специфіку даної галузі їй вдалося досить швидко адаптуватися до умов пандемії та війни і за підсумками десяти місяців 2022 року принести в економіку України 6 млрд доларів США експортної виручки та показати 10% зростання порівняно з попереднім роком. Заробітні плати працівників української ІТ-галузі в середньому в сім разів вищі за середньостатистичну в державі.

Здатність ІТ галузі адаптуватися до роботи у надзвичайних умовах, можливість віддаленої роботи та досить висока заробітна плата фахівців викликають інтерес кандидатів на отримання роботи в ній. З іншого боку, роботодавці, зацікавлені в пошуку висококваліфікованих фахівців, а також рекрутингові агенції, які є посередниками між тими та іншими, мають виявляти основні тенденції, які складаються у цьому сегменті ринку праці, щоб оптимізувати свій пошук.

Традиційно визначальними факторами для визначення рівня заробітної плати працівника вважається рівень його освіти, досвід роботи та уміння працівника втілювати в конкретні справи свої знання та досвід, тобто його психофізичні характеристики. В середовищі ІТ фахівців традиційні вимоги набувають нових відтінків.

В даній роботі за допомогою лінійної регресійної моделі було визначено, що заробітна плата програмістів залежить від загального стажу роботи, досвіду роботи у компанії та від знання англійської мови. Наявність освіти не впливає на зарплату фахівця, проте без неї вкрай важко влаштуватись на роботу. А для отримання більшої заробітної плати ІТ-фахівцю потрібно докладати зусиль, аби отримати оффер в компанію із великою кількістю співробітників. Дискримінація за гендером стосовно заробітної плати існує, проте вона незначна. Результати регресійного аналізу порівнюються з тими, що отримані методом випадкового лісу – одного з

методів машинного навчання. Також було здійснено прогнозування за допомогою регресійних моделей, моделі з використанням random forest та логарифмів. Всі моделі мали схожі показники MAPE та RMSE.

Кластерний аналіз статистичних даних щодо характеристик працівників ІТ галузі підтвердив існування у середовищі ІТ-спеціалістів трьох груп, які можна інтерпретувати як Junior, Middle та Senior спеціалістів. Найменший розкид значень мають Junior розробники, а найбільший – Senior. Абсолютна більшість Senior розробників заробляють 2700-3700\$, Middle - 1250\$-2500\$, а Junior - 800\$-1250\$. А також виявлено залежність між позицією розробника та рівнем знання англійської мови. Переважна більшість Junior спеціалістів мають рівень англійської Intermediate, Middle розробники - Intermediate та Upper-Intermediate, а Senior - Upper-Intermediate та Advanced.

Практична значущість даної роботи полягає у визначенні впливу різних факторів, таких як досвід роботи, навички, освіта та компанія, в якій працює ІТ-фахівець, на його заробітну плату. Це допоможе рекрутинговим агенціям краще зрозуміти ситуацію на ринку праці, а ІТ-спеціалістам – над якими навичками працювати задля збільшення заробітної плати.

Окрім факторів, вплив яких було досліджено, є й інші, які можуть впливати на заробіток в ІТ-секторі. Цими факторами можуть бути місто, в якому працює фахівець, університет, в якому він навчався, сфера компанії, в якій працює робітник, а також стек технологій, якими володіє працівник. Дослідження цих аспектів може дати більш вичерпне уявлення про заробітну плату, яку буде пропонувати компанія даному спеціалісту. При подальших дослідженнях для більш повних та репрезентативних результатів потрібно скористатися вибіркою, що містить більшу кількість показників та спостережень.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. IT Ukraine Assotiation. Do IT like Ukraine 2022. URL: <https://itukraine.org.ua/files/reports/2022/DoITLikeUkraine2022.pdf> (Дата звернення 24.05.2023)
2. Ukrainian Tech Ecosystem Overview. Дашборд технологічної екосистеми України. URL: <https://uatechecosystem.com/dashboard> (Дата звернення 16.05.2023)
3. Солкарян Л. Види ІТ компаній та переваги роботи в них. URL: <https://blog.ithillel.ua/articles/types-of-it-companies-and-their-advantages> (Дата звернення 16.05.2023)
4. Децюра С. Види ІТ-компаній і їх характеристики. URL: <https://i.factor.ua/ukr/journals/nibu/2022/october/issue-80/article-122511.html> (Дата звернення 18.05.2023)
5. Семенишин В. Півтора роки дистанційної роботи. Чому ІТ бізнес досі за? URL: <https://www.epravda.com.ua/columns/2022/01/6/681217/> (Дата звернення 21.05.2023)
6. Бабак А. ІТ-галузь України в 2021 році зросла на 36%, а кількість фахівців збільшилася до 285 тисяч — Ukraine IT Report. URL: <https://dou.ua/forums/topic/36192/> (Дата звернення 21.05.2023)
7. Барвінок В. Ю. Сучасні тренди та проблеми ІТ-сектора в Україні: підготовка та міграція ІТ-фахівців. Механізм регулювання економіки, 2020, № 4. 92-102с.
8. Dsnews.ua Криза? Чи не чули! Як українська ІТ-індустрія зростала в 2020-му і буде рости в новому році. URL: <https://www.dsnews.ua/ukr/economics/pereformatirovanny-rynok-kakoy-budet-ukrainskaya-it-industriya-v-2021-m-posle-pandemii-28122020-410714> (Дата звернення 22.05.2023)

9. А.О. Мельник, В.І. Прима. ІТ-РИНОК В УКРАЇНІ: ТЕНДЕНЦІ РОЗВИТКУ.
URL:
https://er.knutd.edu.ua/bitstream/123456789/18922/1/PIONBUG_2021_P119-120.pdf (Дата звернення 22.05.2023)
10. Російське вторгнення в Україну (з 2022). URL:
[https://uk.wikipedia.org/wiki/Російське_вторгнення_в_Україну_\(з_2022\)](https://uk.wikipedia.org/wiki/Російське_вторгнення_в_Україну_(з_2022))
(Дата звернення 22.05.2023)
11. Дудко В., Мельник Т., Шкальова А., ІТ воєнного часу. Як українська техіндустрія пройшла, можливо, найважчий рік у своїй історії – у шести графіках. URL: <https://forbes.ua/innovations/it-voennogo-chasu-yak-ukrainska-tekhindustriya-proyshla-mozhlivo-nayvazhchiy-rik-u-svoiy-istorii-u-shesti-grafikakh-23122022-10676> (Дата звернення 22.05.2023)
12. Лаущенко О. «До закінчення війни дива не станеться». Чи справді на українському ІТ-ринку криза і з чим уже довелося стикнутися компаніям. URL: <https://dou.ua/lenta/articles/is-there-crisis-nowadays/> (Дата звернення 22.05.2023)
13. Як відключення світла вплинули на ринок праці. URL:
<https://texty.org.ua/fragments/108656/yak-vidklyuchennya-enerhiyi-vplynuly-na-rynok-praci/> (Дата звернення 22.05.2023)
14. Редакція DOU. Ринок праці під час війни: 13% айтивців без роботи, ще половина боїться її втратити. URL: <https://dou.ua/lenta/articles/job-market-during-war-part-1/> (Дата звернення 23.05.2023)
15. Щербина А. Ринок талантів: якою буде частка ІТ у ВВП України. URL:
<https://mind.ua/openmind/20246322-rinok-talantiv-yakoyu-bude-chastka-it-u-vvp-ukrayini> (Дата звернення 22.05.2023)
16. Карпенко Н. В. ІТ-індустрія – локомотив вітчизняної економіки Міжнародна наукова конференція. Київ, АПСВТ, 13-14 квітня 2016 р. 27-28с.
17. Філіпенко Т.Г., Севрук І.М. МІЖНАРОДНА ТОРГІВЛЯ ІТ-ПОСЛУГАМИ: МОЖЛИВОСТІ УКРАЇНИ У СВІТОВОМУ ІНФОРМАЦІЙНОМУ

ПРОСТОPI. URL: <https://conf.ztu.edu.ua/wp-content/uploads/2019/06/84-2.pdf>

(Дата звернення 22.05.2023)

18.Кондіус І.С. ТЕНДЕНЦІЇ РОЗВИТКУ ІТ-РИНКУ В УКРАЇНІ. *ЕКОНОМІЧНИЙ ФОРУМ 1/2019*. 91-99с.

19.Вікулова А.О., Савчук В.В. ПЕРСПЕКТИВИ РОЗВИТКУ РИНКУ ІТ-ПОСЛУГ В УКРАЇНІ. *ЕКОНОМІКА ТА УПРАВЛІННЯ НАЦІОНАЛЬНИМ ГОСПОДАРСТВОМ*, 2020. 27-31с

20.Мельник Т., Завгородня Є. ІТ-СЕКТОР УКРАЇНИ НА СВІТОВОМУ РИНКУ:2022. *ISSN2616-6100. Зовнішня торгівля: економіка, фінанси, право.2022. №6*. 17-36с

21.Ставицький А.В., Навчально-методичний комплекс з курсу „Економетрика”, 2004. 7-26с.

22.Пістунов І. М. , Антонюк О. П., Турчанінова І. Ю. Кластерний аналіз в економетриці. Дніпро, 2008. С 21-38.

23.Матцуга О., Нашлинюк Б. Конструювання ансамблю дерев рішень. алгоритми та їх порівняльний аналіз. *Актуальні проблеми автоматизації та інформаційних технологій*. Том 22, 2018. 62-69с.

24.What is Random Forest URL: <https://www.ibm.com/topics/random-forest> (Дата звернення 15.05.2023)

25.Сергій Волошин. Дані опитувань сайту DOU.ua URL: <https://github.com/devua/csv/tree/master/salaries> (Дата звернення 15.05.2023)

26.Кореляційна матриця. URL: <https://studfile.net/preview/5211482/page:6/> (Дата звернення 15.05.2023)

27.Zach. How to use summary() Function in R (With Examples). URL: <https://www.statology.org/summary-function-in-r/> (Дата звернення 15.05.2023)

28.Available Diagnostic Visualizations . URL: https://docs.tibco.com/pub/spotfire/6.5.2/doc/html/prd/prd_available_diagnostic_visualizations.htm (Дата звернення 15.05.2023)

- 29.Ткаченко О. М., Грійо Тукало О. Ф., Дзісь О. В., Лаховець С. М. МЕТОД КЛАСТЕРИЗАЦІЇ НА ОСНОВІ ПОСЛІДОВНОГО ЗАПУСКУ К-СЕРЕДНІХ З УДОСКОНАЛЕНИМ ВИБОРОМ КАНДИДАТА НА НОВУ ПОЗИЦІЮ ВСТАВКИ. URL: <https://core.ac.uk/download/pdf/52159134.pdf> (Дата звернення 15.05.2023)
- 30.Молчан В., Маринич Т. Моделювання тенденцій на ринку праці ІТ сектору України. Суми, 2020. 27-28с.

ДОДАТОК А

Код на мові програмування R

```
install.packages("factoextra")
install.packages("car")
install.packages("readr")
install.packages("caret")
install.packages("randomForest")
library(randomForest)
library(dplyr)
library(readr)
library(car)
library(ggplot2)
library(factoextra)
library(caret)

salaries = read_csv("/Users/lovinggreen/Downloads/trial_2.csv")
summary(salaries)

#побудова кореляційної матриці
data_for_cor <- data.frame(matrix(nrow = nrow(salaries)))

data_for_cor$salary <- salaries$Salary
data_for_cor$Total_Exp <- salaries$Total_Exp
data_for_cor$Current_Exp <- salaries$Current_Exp
data_for_cor$Age <- salaries$Age
data_for_cor <- subset(data_for_cor, select = -matrix.nrow...nrow.salaries..)
```

```
cor_matrix <- cor(data_for_cor)
```

```
#побудова покрокової лінійної регресії
```

```
min.model <- lm (salaries$Salary ~ 1)
```

```
fwd.model <- step(
```

```
  min.model,
```

```
  direction = "forward",
```

```
  scope = (~ salaries$Total_Exp + salaries$Current_Exp + salaries$Age),
```

```
  trace = 0
```

```
)
```

```
summary(fwd.model)
```

```
#якісний показник статі
```

```
salaries$Gender[salaries$Gender == "Чоловік"] <- "0"
```

```
salaries$Gender[salaries$Gender == "Жінка"] <- "1"
```

```
salaries$Gender <- as.numeric(salaries$Gender)
```

```
cor(salaries$Salary, salaries$Gender)
```

```
m_descrimination <- lm(Salary~Total_Exp+Current_Exp+Gender,  
data=salaries)
```

```
summary(m_descrimination)
```

```
#якісний показник освіти
```

```
salaries$Education[salaries$Education == "Вища"] <- "0"
```

```
salaries$Education[salaries$Education == "Дві вищих"] <- "1"
```

```
salaries$Education[salaries$Education == "Середня"] <- "2"
```

```
salaries$Education <- as.numeric(salaries$Education)
```

```
salaries <- na.omit(salaries)
cor(salaries$Salary, salaries$Education)
```

```
m_education <- lm(Salary~Total_Exp+Current_Exp+Education, data=salaries)
summary(m_education)
```

```
#якісний показник англійська мова
```

```
salaries$English[salaries$English == "Elementary"] <- "0"
salaries$English[salaries$English == "Pre-Intermediate"] <- "1"
salaries$English[salaries$English == "Intermediate"] <- "2"
salaries$English[salaries$English == "Upper-Intermediate"] <- "3"
salaries$English[salaries$English == "Advanced"] <- "4"
salaries$English <- as.numeric(salaries$English)
cor(salaries$Salary, salaries$English)
```

```
m_english <- lm(Salary~Total_Exp+Current_Exp+English, data=salaries)
summary(m_english)
```

```
#якісний показник тип компанії
```

```
salaries$Type_Of_Company[salaries$Type_Of_Company ==
"Аутсорсингова"] <- "0"
salaries$Type_Of_Company[salaries$Type_Of_Company == "Інша"] <- "1"
salaries$Type_Of_Company[salaries$Type_Of_Company == "Стартап"] <- "2"
salaries$Type_Of_Company[salaries$Type_Of_Company == "Продуктова"] <-
"3"
salaries$Type_Of_Company[salaries$Type_Of_Company == "Аутстафінгова"]
<- "4"
salaries$Type_Of_Company <- as.numeric(salaries$Type_Of_Company)
cor(salaries$Salary, salaries$Type_Of_Company)
```

```

m_descrimination <- lm(Salary~Total_Exp+Current_Exp+Type_Of_Company,
data=salaries)
summary(m_descrimination)

#якісний показник кількість співробітників
salaries$Number_Of_Coworkers[salaries$Number_Of_Coworkers == "Лише
я"] <- "0"
salaries$Number_Of_Coworkers[salaries$Number_Of_Coworkers == "до 10
спеціалістів"] <- "1"
salaries$Number_Of_Coworkers[salaries$Number_Of_Coworkers == "до 50"]
<- "2"
salaries$Number_Of_Coworkers[salaries$Number_Of_Coworkers == "до 200"]
<- "3"
salaries$Number_Of_Coworkers[salaries$Number_Of_Coworkers == "до
1000"] <- "4"
salaries$Number_Of_Coworkers[salaries$Number_Of_Coworkers == "понад
1000"] <- "5"
salaries$Number_Of_Coworkers <-
as.numeric(salaries$Number_Of_Coworkers)
cor(salaries$Salary, salaries$Number_Of_Coworkers)

m_coworkers <- lm(Salary~Total_Exp+Current_Exp+Number_Of_Coworkers,
data=salaries)
summary(m_coworkers)

#оцінка моделі

error <- resid(m_english)

```

```
summary (error)
```

```
hist (error)
```

```
outlierTest(m_english)
```

```
par(mfrow = (c(2,2)))
```

```
plot (m_english)
```

```
new_salaries <- salaries[-c(2, 1, 8, 984, 841), ]
```

```
new_m_english <- lm(Salary~Total_Exp+Current_Exp+English,  
data=new_salaries)
```

```
summary(new_m_english)
```

```
qplot(data=salaries, log(Total_Exp), log(Salary))
```

```
model_log <- lm(log(Salary)~log(Total_Exp), data=new_salaries)
```

```
model_log_eng <- lm(log(Salary)~log(Total_Exp)+English, data=new_salaries)
```

```
summary(model_log)
```

```
summary(model_log_eng)
```

```
#кластеризація
```

```
cluster_dataset <- data.frame(matrix(nrow = nrow(salaries)))
```

```
cluster_dataset$salary <- salaries$Salary
```

```
cluster_dataset <- subset(cluster_dataset, select = -matrix.nrow...nrow.salaries..)
```

```
m <- dist(scale(cluster_dataset))
```

```
hc <- hclust(m, method = "ward.D")
plot(hc, cex = 0.9)
```

```
groups3 <- cutree(hc, k = 3)
cluster_dataset$groups3 <- factor(groups3)
```

```
ggplot(data = cluster_dataset, aes(x = "", y = salary)) + geom_boxplot() +
facet_grid(~groups3)

ggplot(data = cluster_dataset, aes(x = "", y = salary, fill = groups3)) +
geom_violin() + facet_grid(~groups3)
```

```
fviz_nbclust(cluster_dataset, kmeans, method = "wss") +
labs(subtitle = "Elbow method") +
geom_vline(xintercept = 3, linetype = 2)
```

```
combined_salaries <- cbind(salaries, cluster_dataset)
combined_salaries <- combined_salaries %>% rename(groups3 = cluster)
```

```
par(mfrow = c(2, 2))
bg1 <- qplot(data = subset(combined_salaries, groups3 == 1), Total_Exp,
Salary, color=factor(English))
bg2 <- qplot(data = subset(combined_salaries, groups3 == 2), Total_Exp,
Salary, color=factor(English))
bg3 <- qplot(data = subset(combined_salaries, groups3 == 3), Total_Exp,
Salary, color=factor(English))
bg1
bg2
```

bg3

```
#randomForest
install.packages("randomForest")
library(randomForest)
tree <- randomForest(Salary ~ Total_Exp+Current_Exp+English, data =
new_salaries, do.trace=50, mtry=3, importance=TRUE, nPerm=3)
tree

in_sample <- createDataPartition(new_salaries$Salary, p=0.75, list=FALSE)
head(in_sample)

salaries_train <- new_salaries[in_sample,]
salaries_test <- new_salaries[-in_sample,]

model_lm <- lm(Salary ~ Total_Exp+Current_Exp+English, data =
new_salaries)
model_rf <- randomForest(Salary ~ Total_Exp+Current_Exp+English, data =
new_salaries)

y <- salaries_test$Salary
yhat_lm <- predict(model_lm, salaries_test)
yhat_rf <- predict(model_rf, salaries_test)
y_log <- log(salaries_test$Salary)
yhat_log <- predict(model_log, salaries_test)
yhat_log_eng <- predict(model_log_eng, salaries_test)

sqrt(sum( (y - yhat_lm)^2)/1244)
sqrt(sum( (y - yhat_rf)^2)/1244)
```

```
sqrt(sum( (exp(y_log) - exp(yhat_log))^2)/1244)
```

```
sqrt(sum( (exp(y_log) - exp(yhat_log_eng))^2)/1244)
```

```
sum(abs((y - yhat_lm) / y) * 100)/1244
```

```
sum(abs((y - yhat_rf) / y) * 100)/1244
```

```
sum(abs((exp(y_log) - exp(yhat_log)) / (exp(y_log)))) * 100)/1244
```

```
sum(abs((exp(y_log) - exp(yhat_log_eng)) / (exp(y_log)))) * 100)/1244
```



Ім'я користувача:
Математичного моделювання та статистики Даценк...

Дата перевірки:
04.06.2023 11:08:11 EEST

Дата звіту:
04.06.2023 11:19:40 EEST

ID перевірки:
1015410365

Тип перевірки:
Doc vs Internet + Library

ID користувача:
100005744



Назва документа: Подрига_Юнькова_плагіат

Кількість сторінок: 54 Кількість слів: 11406 Кількість символів: 86231 Розмір файлу: 3.14 MB ID файлу: 1015073699

5.91% Схожість

Найбільша схожість: 0.85% з джерелом з Бібліотеки (ID файлу: 1015053413)

3.24% Джерела з Інтернету

292

Сторінка 56

4.97% Джерела з Бібліотеки

446

Сторінка 58

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0% Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи

2