Міністерство освіти і науки України

Національний університет «Києво-Могилянська академія»

Факультет інформатики

Кафедра математики

Кваліфікаційна робота

освітній ступінь – бакалавр

на тему: **«ПОБУДОВА МАТЕМАТИЧНОЇ МОДЕЛІ ЗАРОБІТНОЇ ПЛАТИ ТА ПРОГНОЗ НА НАСТУПНИЙ ПЕРІОД»**

Виконала: студентка 4-го року навчання
освітньої програми «Прикладна математика»,
спеціальності 113 Прикладна математика

Дибкалюк Ольга Тарасівна
Керівник: Дрінь С.С.,
кандидат фіз.-мат. наук, ст. викладач

Рецензент_____
(прізвище та ініціали)

Кваліфікаційна робота захищена
з оцінкою _____

Секретар ЕК _____
«____» _____ 20____ р.

Київ 2022

Міністерство освіти і науки України
НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ «КИЄВО-МОГИЛЯНСЬКА АКАДЕМІЯ»
Кафедра математики факультету інформатики

ЗАТВЕРДЖУЮ

Зав. кафедри математики,

проф., д.ф.-м.н.

_____ Б. В. Олійник

(підпис)

«____»_____2021 р.

ІНДИВІДУАЛЬНЕ ЗАВДАННЯ

на кваліфікаційну роботу

студентці 4-го курсу факультету інформатики

Дибкалюк Ользі Тарасівні

Тема: побудова математичної моделі заробітної плати та прогноз на наступний період.

Вихідні дані: Досліджено для побудови та аналізу лінійної авторегресивної моделі.

Зміст ТЧ до кваліфікаційної роботи:

Індивідуальне завдання

Календарний план

Анотація

Вступ

1. Метод дисперсійного аналізу для множинної регресійної моделі зі змінними взаємодій.

2. Основні принципи роботи нелінійної множинної регресії.

3. Побудова та застосування нелінійної множинної регресії для прогнозу даних.

Висновки

Використана література

Додатки

Дата видачі „___” _____ 2022 р Керівник _____*(підпис)* Завдання отримав _____

Тема: П**обудова математичної моделі заробітної плати та прогноз на наступний період.**

Календарний план виконання роботи

Номер етапу	Назва етапів кваліфікаційної роботи	Термін виконання етапів роботи	Примітка
1.	Вибір і обґрунтування теми, постановка проблем і завдань	02.11.2021	
2.	Визначення літератури для роботи.	10.02.2022	
3.	Огляд і аналіз літератури за темою роботи.	15.02.2022	
4.	Оформлення теоретичної частини кваліфікаційної роботи	01.05.2022	
5.	Виконання практичного застосування вибраного методу.	10.05.2022	
6.	Оформлення практичної частини кваліфікаційної роботи	25.05.2022	
7.	Створення презентації кваліфікаційної роботи	25.06.2022	
8.	Захист кваліфікаційної роботи	04.07.2022	

Студент:___Дибкалюк О.Т. _____

Керівник:_____Дрінь С.С._____

“30” червня 2022 р

Зміст

Анотація

У даній роботі описано типову функції заробітків та розглянуто власну модель заробітної плати, а саме нелінійну регресійну логарифмічну модель зі змінними взаємодій категоріальних та бінарних змінних.

У першому розділі даної роботи відбувається дослідження предметної області, описуються методи прогнозування, що можуть бути використані для нашого дослідження, наводиться обґрунтування та опис обраного методу для аналізу та прогнозування.

У другому розділі відбувається вибір виду економіко-математичної моделі, будується обрана модель за нашими даними, обчислюється прогноз по даній моделі та проводиться порівняння з реальними даними. Описано фактори впливу на зарплату, основні проблеми, що виникають під час прогнозування та ключові рішення, що допомагають у створенні моделей. Представлено основні кроки побудови власної моделі та аналіз продуктивності алгоритму. Наведено приклад застосування моделі заробітної плати по Україні із заданими параметрами. Проведено аналіз отриманих результатів.

У висновку описані підсумки дослідження, формуються остаточний звіт з тематики, що вивчається.

Вступ

Після повномасштабного вторгнення росії український IT-ринок ненадовго завмер і лише зараз починає відновлюватися, проте вже у новому вигляді: розробники та рекрутери переорієнтовуються на західний ринок, відгуки на вакансії значно зросли й де-не-де спостерігається спад зарплатних очікувань.

Опитування, проведене українськ DOU, показало, що вже через місяць після початку війни в Україні більшість IT-спеціалістів повернулися до роботи. Переважна більшість компаній повністю відновили та продовжують свою роботу, від 50 до 80% спеціалістів залучені до виконання проектів клієнтів. Завдяки цьому, галузь залишається фінансово стабільною, адже 77% компаній більшою мірою зберегли майже всіх клієнтів та обсяги своїх контрактів. Майже усі IT-бізнеси були готові до такого сценарію – позитивну роль зіграв досвід роботи в умовах пандемії.

Про втрату роботи через війну повідомили усього лишень 6% опитаних. Проте у найближчий місяць частка таких спеціалістів може зрости через оптимізацію бізнесу компаніями.

Також аналітика сайтів для пошуку роботи показує, що сфери IT, продажів, маркетингу, реклами та PR постраждали від наслідків війни найменше. Ба більше, з початку воєнного стану найбільший попит мають фахівці у сфері продажів та IT, інтернет, телекому. Відповідно, для цих фахівців існує ширший вибір пропозицій роботи і вищий рівень зарплат. Сама IT-сфера може стати новим головним рушієм економіки. У її становленні велику роль відіграє політика держави щодо айтівців, наявність чітких сигналів до сфери та грамотна робота зі структурою ринку, що вже змінюється.

Сама Міністерка економіки України Юлія Свириденко зазначила: «Наше завдання – зберегти та підтримати сектор IT, як один з найбільш динамічних та перспективних, що матиме величезне значення після нашої перемоги в процесі відбудови України».

Така обнадійлива статистика та позиція уряду дає можливість прогнозування рівня зарплат – утім, з урахуванням екзогенних факторів. Звісно, неможливо взяти до уваги абсолютно усі ризики і розробити достатню кількість потенційних варіантів розвитку подій – особливо, коли ми говоримо про ймовірність затяжної війни.

З деяких відповідей в опитуванні можна зробити висновок, що найбільш упевнено в довгостроковій перспективі відчувають себе великі гравці, які мають потужні материнські компанії за кордоном. Гравці меншого масштабу говорять про готовність забезпечити стабільну роботу і працюють відповідно до короткострокового плану. [1]

Саме тому аналіз даних та подальше прогнозування є важливою складової діяльності усього ІТ-сектору в Україні.

Об'єкт дослідження: показники ЗП програмістів-розробників ІТ-сфери.

Метою даної роботи є аналіз даних, прогнозування показників та порівняння з реальними даними ЗП програмістів-розробників ІТ-сфери.

Для досягнення поставленої мети необхідно вирішити наступні завдання:

1. Проаналізувати необхідну наукову та навчально-методичну літературу.
2. Проаналізувати діяльність ІТ-сектору, а також вихідні дані.
3. Вивчити методи аналізу та прогнозування.
4. Базуючись на характеристиках даних, обрати найбільш підходящий їм метод аналізу та прогнозування.
5. Провести аналіз даних обраним методом та обчислити прогноз.

Методи дослідження: статистичний аналіз, методи прогнозування.

Розділ 1. Аналіз даних та вибір методу їх прогнозування

1.1 Характеристика вихідних даних

Компанія DOU.ua з'явилася у 2005 році як блог на WordPress. Сьогодні на DOU більше 400 000 користувачів та 9 млн. переглядів сторінок на місяць. Щомісяця у нас розміщується близько 6000 IT-вакансій від 1500 компаній.

Для нашого дослідження ми взяли результати опитування, що проводяться компанією регулярно, починаючи з 2011 року – кожного грудня та червня.

Дані складаються з певної кількості стовпців: різноманітних характеристик айтівця та його зарплати.

Для зручного користування даними була проведена наступна їхня обробка. З усіх позицій було обрано software engineer з різним досвідом. Дані представлені за червень 2021 року, усього у вибірці було представлено 5646 респондентів.

1.2 Дисперсійний аналіз (ANOVA)

Спершу введемо поняття множинної лінійної регресії.

Позначимо y залежною (досліджуваною, або пояснюваною) змінною, яка лінійно пов’язана з k незалежними (пояснювальними) x за допомогою параметрів β_k.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \qquad (1.2.1)$$

Формула (1.1) є визначенням багатофакторної (множинної) лінійної регресії. Параметри $\beta_1, \beta_2, \ldots, \beta_k$, називаються регресійними коефіцієнтами, які відповідно пов’язані з $x_1, x_2, \ldots, x_k$ та ε – випадкова похибка, що відображає різницю між спостережуваною та прогнозованою лінійною залежністю. Обґрунтувань її появи є кілька – це може бути спільний ефект тих змінних, які не включені в модель, випадкові фактори, що не враховуються у моделі, екзогенні фактори, людський фактор тощо.

β_0 – параметр, який не відповідає за жодний з факторів, називають константою. Формально це є значення функції при нульовому значенні всіх факторів. В аналітиці прийнято вважати, що константа - це параметр при «факторі», рівному 1. Лінійна модель може існувати як з константою, так і без неї.

i-коефіцієнт регресії β_i показує швидкість зміни функції y при зміні аргументу x_i на одиницю.

Припускаючи, що $E(\varepsilon) = 0$,

$$\beta_i = \frac{\partial E(y)}{\partial x_i}$$

[16]

ANOVA для множинної лінійної регресії

Дисперсійний аналіз (англ. analysis of variance (ANOVA)) – статистичний метод аналізу результатів, що залежать від якісних ознак.

Кожен фактор може бути дискретною чи неперервною випадковою змінною, яку розділяють на декілька сталих рівнів (градацій, інтервалів). Якщо кількість вимірювань на всіх рівнях кожного з факторів однакова, то дисперсійний аналіз називають рівномірним, інакше — нерівномірним.

Дисперсійний аналіз базується на наступному принципі: якщо на випадкову величину діють взаємно незалежні фактори A, B, …, то загальна дисперсія дорівнює сумі дисперсій, зумовлених дією окремо кожного з факторів:

$$\sigma^2 = \sigma_A^2 + \sigma_B^2 + \cdots$$

Множинна лінійна регресія намагається підігнати лінію регресії для залежної змінної, що має більше ніж одну пояснювальну змінну. Розрахунки ANOVA для множинної регресії майже ідентичні розрахункам для простої лінійної регресії, за винятком того, що ступені свободи коригуються для відображення кількості пояснювальних змінних, що включені в модель.

Для p пояснювальних змінних, ступені свободи (DFM) моделі дорівнюють p, а помилка ступенів свободи (DFE) дорівнює $(n - p - 1)$. Сумарні ступені свободи (DFT) дорівнюють $(n - 1)$, сумі DFM і DFE.

Прийняті скорочення:

SST – Sum of Squares Total;

SSE – Sum of Squares Error;

SSM – Sum of Squares Model.

Відповідна таблиця ANOVA наведена нижче (табл. 1.2.1)

Джерело	Ступені свободи	Сума квадратів	Середнє квадратичне	F
Модель	p	$\sum(\hat{y}_i - \bar{y})^2$	SSM/DFM	MSM/MSE
Помилка	$n - p - 1$	$\sum(y_i - \hat{y}_i)^2$	SSE/DFE	
Сумарно	$n - 1$	$\sum(y_i - \bar{y})^2$	SST/DFT	

У множинній регресії тестова статистика MSM/MSE має $F(p, n - p - 1)$ розподіл.

Нульова гіпотеза стверджує, що 1 =$\beta = \beta_2 = ... = \beta_P = 0$, а альтернативна гіпотеза стверджує, що принаймні один з параметрів $\beta j \neq 0, j = 1,2,3 ... p$

Великі значення тестової статистики надають докази проти нульової гіпотези.

Відношення SSM/SST = R^2 відоме як квадратичний коефіцієнт множинної кореляції.

Через те, що наші дані містять і якісні, і неперервні змінні, для їхнього аналізу було запропоновано використати саме цей метод ANOVA. [16]

1.3 Варіанти регресії. Інтерпретація коефіцієнтів у логарифмічних моделях з логарифмічними трансформаціями

Розглядаючи просту двовимірну лінійну модель $Yi = \alpha + \beta Xi + \varepsilon_i$,

доцільним буде пройтися по чотирьох можливих комбінаціях перетворень із логарифмами:

- Лінійна форма без перетворень;
- лінійно-логарифмічна модель;
- логарифмічно-лінійна модель;
- логарифмічно-логарифмічна модель.

	X	
Y	X	log$X$
Y	*linear* $\hat{Y}_i = \alpha + \beta X_i$	*linear-log* $\hat{Y}_i = \alpha + \beta \log X_i$
log$Y$	*log-linear* $\log\hat{Y}_i = \alpha + \beta X_i$	*log-log* $\log\hat{Y}_i = \alpha + \beta \log X_i$

Логарифмічне перетворення змінних у регресійній моделі є дуже поширеним способом обробки ситуацій, коли між незалежною та залежною змінними існує нелінійний зв'язок. Використання логарифма однієї або кількох змінних замість початкової форми робить ефективним нелінійне відношення, зберігаючи при цьому лінійну модель.

Логарифмічні перетворення також є зручним засобом перетворення сильно скошеної змінної в більш схожу на нормальну.

Нагадаємо, що в моделі лінійної регресії $lgYi = \alpha + \beta Xi + \varepsilon_i$, коефіцієнт β дає нам пряму зміну Y для зміни X на одну одиницю. Додаткова

інтерпретація не потрібна. Буквальна інтерпретація все ще зберігатиметься, навіть коли змінні були логарифмічно перетворені, але після модифікації зазвичай має сенс інтерпретація змін не в логарифмічних одиницях, а у відсотках.

Кожна логарифмічно перетворена модель розглядається по черзі нижче (за винятком перщої)

a) **Лінійно-логарифмічна модель:** $\boldsymbol{Yi = \alpha + \beta lgXi + \varepsilon i}$

У лінійно-логарифмічній моделі буквальна інтерпретація оціненого коефіцієнта $\hat{\beta}$ полягає в тому, що збільшення log X на одну одиницю призведе до очікуваного збільшення Y на $\hat{\beta}$ одиниць. Щоб побачити, що це означає для X, ми можемо використати прописати наступну рівність:

$$log\, X + 1 = log\, X + log\, e = log(eX)$$

, що була отримана за допомогою властивостей логарифмів і експоненційних функцій.

Іншими словами, додавання 1 до log X означає множення самого X на e ≈ 2,72.

Таку пропорційну зміну можна перетворити на зміну у відсотках, віднімаючи 1 і помножуючи на 100. Отже, інший спосіб сказати «множення X на 2,72» — це сказати, що X збільшується на 172% (оскільки 100 × (2,72 − 1) = 172).

Отже, з точки зору зміни X:

• $\hat{\beta}$ – очікувана зміна Y, якщо X помножити на e.

• $\hat{\beta}$ – очікувана зміна Y, коли X збільшується на 172%

• Для інших відсоткових змін у X ми можемо використовувати такий результат: Очікувану зміну в Y, пов'язану із $p\%$ збільшенням X, можна розрахувати як $\hat{\beta} \cdot log([100 + p]/100)$. Отже, щоб визначити очікувану зміну, пов’язану зі збільшенням X на 10%, треба помножити $\hat{\beta}$ на $log(110/100)$ = log(1,1) = 0,095. Іншими словами, $0{,}095\hat{\beta}$ є очікуваною зміною Y, коли X множиться на 1,1, тобто збільшується на 10%.

• Для малих p приблизно $log([100 + p]/100) \approx p/100$. Для $p = 1$ це означає, що $\hat{\beta}$ /100 можна приблизно інтерпретувати як очікуване збільшення Y від збільшення X на 1%.

b) **Логарифмічно-лінійна модель:** $\boldsymbol{lgYi = \alpha + \beta Xi + \varepsilon i}$

У логарифмічно-лінійній моделі буквальна інтерпретація оціненого коефіцієнта $\hat{\beta}$ полягає в тому, що одна одиниця збільшення X призведе до очікуваного збільшення log Y на $\hat{\beta}$ одиниць. З точки зору самого Y, це означає, що очікуване значення Y множиться на $e^{\hat{\beta}}$. Отже, з точки зору впливу змін X на Y:

• Кожне збільшення X на 1 одиницю множить очікуване значення Y на $e^{\hat{\beta}}$.

• Щоб обчислити вплив на Y іншої зміни X, ніж збільшення на одну одиницю, нам потрібно включити c до показника ступеня. Ефект збільшення с −одиниці X полягає в тому, щоб помножити очікуване значення Y на $e^{c\hat{\beta}}$

Таким чином, ефект для збільшення X на 5 одиниць буде дорівнювати $e^{5\hat{\beta}}$

• Для малих значень $\hat{\beta}$ приблизно $e^{\hat{\beta}} \approx 1+\hat{\beta}$. Ми можемо використовувати це для наступного наближення швидкої інтерпретації коефіцієнтів: $100\times \hat{\beta}$ – очікувана зміна у відсотках в Y для збільшення X на одну одиницю. Наприклад, для $\hat{\beta} = .06$, $e^{.06} \approx 1.06$, отже, зміна X на 1 одиницю

відповідає (приблизно) очікуваному зростанню Y на 6%.

c) **Модель Log-log:** $\boldsymbol{logYi = \alpha + \beta logXi + \varepsilon i}$

У випадках, коли як залежна змінна, так і незалежна змінна(і) перетворені в логарифмічну форму, інтерпретація є комбінацію наведених вище випадків лінійного логарифмічного та логарифмічного лінійного. Інакше кажучи, інтерпретація надається як очікувана відсоткова зміна Y, коли X збільшується на певний відсоток. Такі відношення, де як Y, так і X логарифмічні, зазвичай називаються еластичними в econometriці, а коефіцієнт log X називають **еластичністю.**

Отже, з точки зору впливу змін у X на Y:

• множення X на e помножить очікуване значення Y на $e^{\hat{\beta}}$

• Щоб отримати пропорційну зміну Y, пов'язану з p −відсотковим збільшенням X, треба обчислити $a = log([100 + p]/100)$ і взяти $e^{a\hat{\beta}}$

[17]

Еласти́чність попиту, коефіцієнт еластичності — вказує відносну зміну одного економічного показника за одиничної відносної зміни іншого показника, його детермінанта; відношення відсоткової зміни одного показника (функції %) до відсоткової зміни іншого показника (аргументу %)

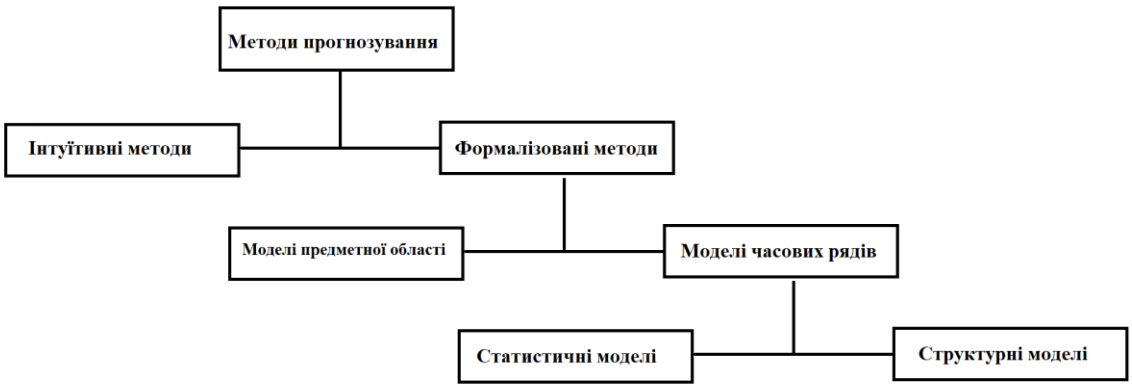
$$E_{x(y)} = \frac{\frac{\Delta Y}{Y}}{\frac{\Delta X}{X}}$$

, де Ex(y) – коефіцієнт еластичності, $\Delta Y -$ зміна функції $Y = f(x)$, $\Delta X -$зміна аргументу X.[18]

1.4 Методи прогнозування

У загальному розумінні прогнозування (грец.: знання наперед) – це вид пізнавальної діяльності людини, спрямованої на формування науково обґрунтованого можливого стану об’єкта у майбутньому, а також альтернативних шляхів і строків досягнення такого стану на основі всебічного аналізу і вивчення тенденцій його розвитку в умовах дії об’єктивних законів природи. Наявність прогнозу дає можливість уникнути помилок, упереджених або запізнілих рішень, попередити несприятливі та небажані події [2].

Нижче наведені види методів та моделі прогнозування



(рис. 1.4.1).

Щоб дати необхідну теоретичну основу для дослідження, дано коротку характеристику кожного з представлених видів прогнозів.

Суть інтуїтивного методу прогнозування полягає у проведенні експертами інтуїтивного та логічного аналізу проблеми, що дозволяє отримати правильне рішення. В основі формалізованих методів прогнозування лежить побудова прогнозів формальними засобами математичної теорії, що значно підвищує достовірність та точність прогнозів, при цьому скоротивши час виконання операції. Моделі предметної області є математичними моделями

прогнозування, для побудови яких використовуються закони предметної галузі. Моделі часових рядів є математичними моделями прогнозування, що встановлюють, чи є залежність між історичними значеннями всередині самого процесу, щоб надалі обчислити прогноз майбутніх значень.

Дані моделі універсальні для сфер різного типу, тому що їхній загальний вигляд не залежить від природи тимчасового ряду. У статистичних моделях визначення залежності між майбутніми і минулими значеннями представляється у вигляді рівняння.

Статистичні моделі включають:

1) моделі регресії (лінійної та нелінійної);

2) авторегресійної моделі;

3) моделі з експонентним згладжуванням;

4) модель, що будується за вибіркою максимальної схожості.

У структурних моделях визначення залежності майбутніх значень від попередніх значень представляється у вигляді структури. Найчастіше під цим мається на увазі наявність деяких правил переходів.

Структурний тип моделі складається з:

1) моделей нейронних мереж;

2) моделей, що ґрунтуються на використанні ланцюга Маркова;

3) моделей, які використовують класифікаційні та регресійні дерева.

Основним завданням є отримання інформації про можливі стани даних зарплат у майбутньому та подальших шляхах досягнення цих цілей. У такому разі найкращим методом прогнозування буде регресійний аналіз (статистичний метод)[3].

Статистичний метод використовується з метою визначення залежності деякої величини від іншої величини, а може і кількох інших величин, що дозволяє встановити зміни у середовищі урахування впливів цих змін на досліджуваний показник. Оскільки дані зарплатні є економічними показниками, отже, прогноз щодо них буде більш повним, якщо використовувати економічні дослідження[4]. Дані дослідження найкраще проводити, використовуючи методи такої науки, як економетрика.

Економетрика – це наука, що вивчає кількісні та якісні економічні взаємозв'язки за допомогою статистичних та інших математичних методів та моделей. [6] Для того, щоб вибрати найбільш підходящий економетричний метод аналізу та прогнозування, необхідно розглянути всі методи економетричного дослідження.

1.4.1 Основні економетричні методи прогнозування

Значимість економетрики полягає в застосування її методів. Воно дозволяє виявити суттєві зв'язки між явищами, а також дати прогноз подальшого розвитку даного явища, оцінити та перевірити економічні наслідки, що виникли при прийнятті будь-яких важливих рішень управління. Щоб вибрати найбільш підходящий метод економетричного дослідження, необхідно розглянути та описати основні економетричні методи [7].

1. Першим таким методом є *парний регресійний аналіз*.

Між економічними змінними існують лише статистичні та кореляційні залежності. Виражаючи змінні через X, Y, отримаємо залежність такого виду

Ця залежність матиме назву функції регресії Y на X. X має назву незалежної змінної, що пояснює залежну, Y – залежна змінна або пояснювана. У випадку розглядання двох випадкових величин використовують парну регресію.

2. *Нелінійна регресія.*

Не всі економічні залежності є лінійними, тому їх моделювання лінійними рівняннями регресії не є доцільним. Задля цього використовують такий вид регресійного аналізу, як нелінійна регресія, що в ньому експериментальні дані моделюються за допомогою функції. Ця функція є нелінійною комбінацією параметрів моделі і залежить від однієї та більше незалежних змінних. Дані апроксимуються методом послідовних наближень [8].

3. *Множинна регресія*.

Для вирішення таких завдань, як розв’язання проблем попиту, вивчення функцій витрат виробництва, прибутковість акцій тощо, у макроекономічних розрахунках застосовують множинну регресію. Рівняння множини представляється у вигляді:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon \quad ,$$

Де y є залежною (досліджуваною, або пояснюваною) змінною, що лінійно пов’язана з k незалежними (пояснювальними) x за допомогою параметрів β_k. Параметри $\beta_1, \beta_2, \ldots, \beta_k$, називаються регресійними коефіцієнтами, які

відповідно пов’язані з $x_1, x_2, \ldots, x_k$ та ε – випадкова похибка, що відображає різницю між спостережуваною та прогнозованою лінійною залежністю.

β_0 – параметр, який не відповідає за жодний з факторів, називають константою. Формально це є значення функції при нульовому значенні всіх факторів. В аналітиці прийнято вважати, що константа - це параметр при «факторі», рівному 1. Лінійна модель може існувати як з константою, так і без неї. [9].

4. *Фіктивні змінні.*

Якщо при побудові моделі ми маємо справу з якісними змінними, їх необхідно перевести в кількісні. Змінні, що мають таку конструкцію, носять назву фіктивних змінних, що кількісним чином описують якісні ознаки.

5. *Системи економічних рівнянь.*

Використання системи рівнянь часто зумовлюється тим, що деякі процеси моделюються відразу декількома рівняннями, які містять власні та повторювані змінні одночасно. [10].

6. *Часові ряди.*

Даний метод представляється у вигляді сукупності значень якогось показника протягом деяких моментів часу, що йдуть один за одним. Кожен рівень y_t формується тоді, коли на нього впливають короткочасні, тривалі та випадкові фактори. Найбільш значний вплив на явище надають тривалі та постійні фактори. Дані фактори формують основну тенденцію ряду – тренд T_t. Такі фактори, як короткочасні та періодичні, утворюють формування сезонних коливань ряду S_t. Випадкові чинники – це відображення випадкових змін рівнів ряду. [11]

1.4.2 Вибір економетричного методу прогнозування

1. Часовий ряд

Дані показників заробітної плати є набором числових даних протягом послідовних періодів часу. Для даних такого типу найбільше підходить метод аналізу часових рядів, за допомогою якого можна передбачити значення числової змінної, спираючись на її минулі та справжні значення. Основу аналізу часових рядів можна описати наступним припущенням: фактори, що впливають на об'єкт дослідження в сьогодення та минуле, будуть на нього впливати і в майбутньому. Отже, основні завдання аналізу часових рядів полягають у виділенні та ідентифікації факторів, що мають

значення для прогнозування. Щоб вирішити це завдання, було розроблено багато математичних моделей, які призначені для дослідження коливань компонентів, які входять у модель часового ряду [11]. Кожен рівень часового ряду, як було зазначено раніше, формується під впливом короткочасних, тривалих та випадкових факторів.

Основні визначення, що формуються факторами, описаними вище:

- T_t – Основна тенденція ряду;
- S_t – Сезонні коливання ряду;
- C_t – Циклічна компонента;
- E_t – Випадкові фактори.

Рівні часового ряду можна виразити у вигляді суми систематичної, яка також може бути регулярною або детермінованою та випадковою, яка не залежна від часу, складових. Отримати регулярну складову можна, склавши тренд, циклічну та сезонну компоненти. Утім, бувають випадки, коли ця складова може і не включати всі три компоненти одночасно.

Тим не менш, для нашого випадку аналіз часових рядів не є найбільш підходящою версією, адже ми оцінюємо не одне значення протягом певного періоду часу, а цілу базу даних по різних людях. Отже, метод часових рядів можна сміливо відхиляти.

2. Нелінійна регресія.

Відповідно до економічної теорії, індивідуальна норма (ставка) заробітної плати визначається як цінність ринкових товарів, які можуть бути куплені за гроші, зароблені за годину праці. Дані за нормами заробітної плати, що збираються по домашніх господарствах, зазвичай відповідають цьому теоретичному становищу зі значною помилкою. [7]

Економетрична література за факторами, що визначають заробітну плату, здебільшого описується рівнянням нелінійної регресії виду:

$$\ln y_i = f(s_i, x_i, z_i) + u_i, i = 1,..n \quad (1.4.1)$$

де $\ln y_i$ - натуральний логарифм заробітків або заробітної плати для i-го індивідуума;

s_i - рівень освіти чи освітніх досягнень;

x_i – внесок професійного досвіду у людський капітал;

z_i - інші фактори, що впливають на заробіток (раса, місце проживання людини, стать тощо);

u_i - випадковий залишок, що відображає вплив неспостережуваних характеристик здібностей і внутрішню стохастичність значень заробітків, що спостерігаються.

Як правило, передбачається, що u_i розподілена нормально з середнім нуль та постійною дисперсією. [7]

Рівняння (1.3.1) часто називають статистичною функцією заробітків.

$$\ln Y_s = \ln Y_0 + rs + u \text{ (1.4.2)}$$

Рівність (1.3.2) - це найбільш загальна форма функції заробітків, де

$\ln Y_0 -$ логарифмований заробіток без урахування досвіду (оцінюваний вільний член) та подальшого професійного розвитку, $rs -$ взаємодія норми віддачі на навчання та кількості років освіти (оцінений коефіцієнт нахилу).

$$\beta_0 = \ln Y_0$$

Сам заробіток можна легко знайти за допомогою оберненої функції:

$$Y_0 = e^{\beta_0}$$

Ця проста специфікація функції заробітків була виведена Дж. Мінцером (Jacob Mincer, 1974), коли він включив до рівня моделі ефект впливу загальних форм підвищення кваліфікації. У цьому випадку рівняння (1.4.2) зводиться до виду

$$\ln Y_i = \beta_0 + \beta_1 s_i + \beta_2 k_i X_i + u_i \text{ (1.4.3)}$$

Де β_1 – норма результату класичного навчання;

β_2 - норма віддачі від навчання на роботі;

k_i - частка і-го інтервалу часу, витрачена на загальне підвищення кваліфікації; $X_i -$ тривалість трудового стажу і –го працівника, досягнута до і-го інтервалу часу. На жаль, дані k_i зазвичай недоступні.

У рамках теорії людського капіталу ми припускаємо, що заробітки не будуть сталими після закінчення школи, а набудуть параболічної форми з піком десь посередині. Дану гіпотезу можна імплементувати в наше рівняння наступним

чном: його до лінійного вигляду зі шкільного навчання, але квадратичного за стажем

$$\ln Y_i = \beta_0 + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + u_i \text{ (1.4.4)}$$

Якщо функція заробітків під впливом показників трудового стажу набуває увігнутої форми, як це передбачає теорія людського капіталу, тоді оцінки β_2 повинні бути позитивними, а оцінки — β_3 негативними.

Очевидно, що якщо здібності корелюють з роками навчання і якщо індивідууми з кращою освітою отримують більше можливостей професійного росту на роботі, тоді структура доходів більш розвинених після отримання освіти буде вище, ніж у менш освічених. Один із способів урахування цього ефекту – відобразити у рівнянні взаємодію факторів навчання та стажу. Так, узагальнивши рівність (1.4.1), отримаємо:

$$\ln Y_i = \beta_0 + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 s_i X_i + u_i \text{ (1.4.5)}$$

Де вплив професійного досвіду на логарифм заробітку, $\partial \ln Y_i / \partial X_i = \beta_2 + 2\beta_3 X_i + \beta_4 s_i$, залежить від рівня професійного досвіду X_i і від рівня освіти s_i.

В емпіричних дослідженнях факторів, що визначають зароблену плату, часто використовуються фіктивні змінні. Припустимо, що є підгрупа осіб, що у неї з якихось причин рівень заробітків на постійну величину у відсотках нижчий, ніж у решти населення (незалежно від освіти та професійного досвіду). Звідси випливає, що логарифм заробітків цих людей відрізняється на постійну величину.

Функцію заробітків (1.4.5) можна змінити з урахуванням цього ефекту для осіб вище вказаної категорії за допомогою введення фіктивної змінної C_{1i}. Вона набуває значення 1, якщо індивідуум належить до першої категорії, або 0 - у протилежному випадку. Ця фіктивна змінна додається рівняння (1.4.4), що призводить до співвідношення:

$$\ln Y_i = \beta_0 + \alpha_1 C_{1i} + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 s_i X_i + u_i \text{ (1.4.6)}$$

де α_1— постійна різниця в логарифмах заробітків людей з категорії та іншим населенням, незалежно від рівня навчання або професійного досвіду.

Варто зазначити, що в термінах відсоткових змін $\alpha_1 = ln(1 + d_1)$, де d_1— відсоткова зміна в характеристиках заробітку осіб даної категорії; для невеликих значень, близьких до нуля, $\alpha_1 \approx d_1$, але якщо α_1більше, ніж 0,15-0,20, то для отримання оцінки 4 необхідно потенціювання співвідношення, що зв'язує α_1та d_1.

Ця специфікація може бути розширена - наприклад, на стать. Уявімо, що індивідууми розділені на чоловіків та жінок, і висловлена гіпотеза про те, що заробітки чоловіків відрізняються від заробітків жінок, але при цьому освіта та стаж однакові для всіх індивідуумів. Введемо фіктивну змінну для статі D_{Gi}, що приймає значення 1 для жінок і 0 для чоловіків. Специфікація у вигляді рівняння з фіктивними змінними, враховуючи категорію та стать, виглядатиме так:

$$\ln Y_i = \beta_0 + \alpha_1 C_{1i} + \alpha_G D_{Gi} + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 s_i X_i + u_i$$

(1.4.7)

де α_G інтерпретується як різниця логарифмів заробітків для жінок і чоловіків, незалежно від того, до якої з категорій – 1 та 0 – відноситься людина (враховуймо однаковий рівень досвіду та освіти).

Подальший розвиток специфікації з фіктивними змінними рівняння (4.6) полягає в розгляданні різноманітних варіантів взаємодії. Припустимо, наприклад, що висунута гіпотеза про відмінність заробітків чоловіків та жінок, що належать до категорії 1 (що еквівалентно припущенню про залежність логарифму заробітку індивідуумів першої групи від статі). Процедура включення цієї взаємодії між категоріями і статтю полягає у тому, що ми вводимо фіктивну змінну $D_{G1,i}$ як добуток C_{1i} і D_{Gi}.

Звідси, $D_{G1,i} = C_{1i} \times D_{Gi}, i = 1, \dots n$

Додаємо цю фіктивну змінну взаємодії до рівняння (1.4.7). В результаті отримуємо

$$\ln Y_i = \beta_0 + \alpha_1 C_{1i} + \alpha_G D_{Gi} + \alpha_{G1} D_{G1,i} + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 s_i X_i + u_i$$

(1.4.8)

Розглянемо інтерпретацію параметрів такої специфікації. Припустимо, що змінні освіти та досвіду дорівнюють нулю.

Відповідно, тоді очікуваний логарифм доходів для чоловіків у категорії 0 дорівнює β_0; оскільки

$$\beta_0 + \alpha_1 C_{1i} + \alpha_G D_{Gi} + \alpha_{G1} D_{G1,i} = \beta_0 + \alpha_1 \times 0 + \alpha_G \times 0 + \alpha_{G1} \times 0 = \beta_0$$

($D_{G1,i} = C_{1i} \times D_{Gi} = 0 \times 0 = 0$).

Для жінок у категорії 0 дорівнює $\beta_0 + \alpha_G$;

$$\beta_0 + \alpha_1 C_{1i} + \alpha_G D_{Gi} + \alpha_{G1} D_{G1,i} = \beta_0 + \alpha_1 \times 0 + \alpha_G \times 1 + \alpha_{G1} \times 0 = \beta_0 + \alpha_G$$

$(D_{G1,i} = \mathrm{C}_{1i} \times D_{Gi} = 1 \times 0 = 0)$.

Для чоловіків у категорії дорівнює 1 дорівнює $\beta_0 + \alpha_1$;

$\beta_0 + \alpha_1 \mathrm{C}_{1i} + \alpha_G D_{Gi} + \alpha_{G1} D_{G1,i} = \beta_0 + \alpha_1 \times 1 + \alpha_G \times 0 + \alpha_{G1} \times 0 = \beta_0 + \alpha_1$

$(D_{G1,i} = \mathrm{C}_{1i} \times D_{Gi} = 0 \times 1 = 0)$.

Для жінок у категорії 1 дорівнює $\beta_0 + \alpha_1 + \alpha_G + \alpha_{G1}$;

$\beta_0 + \alpha_1 \mathrm{C}_{1i} + \alpha_G D_{Gi} + \alpha_{G1} D_{G1,i} = \beta_0 + \alpha_1 \times 1 + \alpha_G \times 1 + \alpha_{G1} \times 1 = \beta_0 + \alpha_1 + \alpha_G + \alpha_{G1}$

$(D_{G1,i} = \mathrm{C}_{1i} \times D_{Gi} = 1 \times 1 = 1)$.

Отже, у такій специфікації вплив статевої ознаки на очікуваний логарифм заробітків залежить від того, чи належить людина до категорії 0 або 1, або від конкретної статі індивідууму в категорії 1. Нульова гіпотеза про те, що ефект взаємодії дорівнює нулю, відповідає перевірці рівності ($\alpha_{G1} = 0$).

$$\alpha_{G1} D_{G1,i}$$

Двійкові змінні з кількома категоріями

Припустимо, що ми хочемо зараз проаналізувати, чи є відмінності у заробітній платі у айтівців одного рівня в різних регіонах. Регіон проживання також є якісним фактором, проте, на відміну від статі, що може приймати лише 2 значення, країни, як правило, розподілені на більш ніж 2 регіони.

Уявімо, що в країні є J регіонів. Розглянемо в якості початкової області один з них. Назвемо його областю 1 і визначимо $J - 1$ фіктивні змінні наступним чином:

$$R_1 = \begin{cases} 1, \text{якщо індивідуум живе в регіоні 1} \\ 0, \text{якщо індивідуум не живе в регіоні 1} \end{cases}$$

.

.

$$R_J = \begin{cases} 1, \text{якщо індивідуум живе в регіоні } J \\ 0, \text{якщо індивідуум не живе в регіоні } J \end{cases}$$

Щойно ми визначили фіктивні змінні, можемо розглянути модель

$$\ln Y_i = \beta_0 + \delta_1 R_1 + \cdots + \delta_j R_j + \alpha_1 C_{1i} + \alpha_G D_{Gi} + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 s_i X_i + u_i$$

(1.4.9)

δ_1 відображає різницю в заробітній платі між тими особами, які проживають у регіоні 2, та особами, які проживають в регіоні 1.

…

δ_j фіксує різницю в заробітній платі між тими особами, які проживають в регіоні J, і тими, хто проживає в регіоні 1.

Фінальна запропонована модель може бути записана таким чином:

$$\ln Y_i = \beta_0 + \delta_2 R_2 + \cdots + \delta_j R_j + \alpha_1 C_{1i} + \alpha_G D_{Gi} + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 s_i X_i + u_i$$ (1.4.10)

1.4.3 Використання методу дерева рішень для пошуку глибокої взаємодії факторів

В попередньому розділі ми розглянули нововведенні змінні, що відповідають за взаємодію між двома або більшою кількістю факторів. Така змінна може мати більший вплив на модель, аніж кожен фактор окремо. В регресії з великою кількістю категоріальних та двійкових змінних таких комбінацій може бути безліч – починаючи від першого рівня (з двома множниками) і закінчуючи складними – 5+ взаємодій. Традиційний спосіб дослідження взаємодій спирається на вимірювання різної важливості. Однак ці способи не дають уявлення про взаємодію другого чи третього порядку. Ідентифікація цих інтеракцій важлива для створення кращих моделей, особливо під час пошуку функцій для використання в лінійних моделях. Для того, аби не перевіряти кожну інтеракцію вручну і пробувати щоразу нову модель, можна використати метод пошуку таких взаємодій, який базується на дереві рішень.

Xgbfi — це аналізатор дампу моделі XGBoost, який ранжирує функції, а також взаємодію функцій за різними показниками, направлений на виявлення взаємодій вищого роду.

XGBoost – це оптимізована розподілена бібліотека підвищення градієнта, розроблена для високої ефективності, гнучості та портативності. Він реалізує алгоритми машинного навчання під фреймворком Gradient Boosting. XGBoost забезпечує паралельне збільшення дерева (також відоме як GBDT, GBM), яке швидко і точно вирішує багато проблем науки про дані.

Методи, засновані на деревах, чудово використовують взаємодію функцій або змінних. У цьому розділі представлені окремі дерева регресії та boosting методу ансамблю (поєднання дерев), що можуть бути використані для моделювання нелінійних відносин і ефектів взаємодії.

Одиночні дерева регресії

При моделюванні функції заробітків необхідно зробити припущення щодо її функціональної форми. На відміну від загальноприйнятої форми лінійної функції, дерева регресії мають зовсім інший підхід. Коли для моделювання функції використовується дерево регресії, функціональна форма не застосовується, проте взаємодія між змінними дозволяється.

Розглядаючи модель регресії, класична функціональна форма виглядає наступним чином:

$$f(X) = \beta_0 + \sum_{\rho}^{j-1} X_j \beta_j \text{ (1.4.11)}$$

; де X – $(n \times p)$-матриця предикторів, n - кількість спостережень і p прогнозів. З іншого боку, дерева регресії припускають модель вигляду:

$f(X) = \sum_{M}^{m-1} C_m \, 1_{(X \in R_m)}$; (1.4.12)

де $R_1; \ldots; R_m$ представляють розбиття простору незалежних змінних, а C_m – середнє значення всіх спостережень, що належать до області R_m.

Алгоритм для побудови одиночних дерев регресії (Алгоритм 1.4.1):

a) Використовуйте рекурсивне розбиття, щоб розділити простір предикторів на набір можливих значень для компонентів $X_1, X_2, \ldots X_p$ на M різних областей, що не накладаються - $R_1; R_2; \ldots; R_M$. Середнє спостережень у межах m −ої області вибираються для того, щоб мінімізувати суму квадратів залишків RSS=$\sum_{m=1}^{M} \sum_{i \in R_m} (r_i)^2$, де залишки обчислюються як $r_i = y_{im} - \hat{y}_{R_m}$.
b) Для кожного спостереження, яке потрапляє в область R_M, зробіть прогноз рівним $\hat{y}_{R_m}$.

Алгоритм 1.3.3.1 описує два кроки, необхідні для побудови одиничних дерев регресії.

Дерева регресії мають кілька переваг: вони можуть легко моделювати взаємодії між коваріатами, вони можуть легко обробляти категоріальні коваріати та справляються з відсутними дані. При моделюванні EPF дерева регресії ідеально підходять для врахування ієрархічної структури освітніх

систем, що характеризується взаємодією між (і всередині) різними рівнями. Незважаючи на переваги з точки зору додаткової гнучкості, дерева регресії мають і деякі недоліки: їм, як правило, притаманна висока дисперсія. Окрім того, вони є чутливими до викидів. Однак існують методи, які шляхом агрегування інформації з багатьох дерев в ансамблі (наприклад, boosting) значно покращують точність прогнозу.

Ансамбль регресійних дерев: boosting

Boosting – це поетапна процедура, що об'єднує інформацію з багатьох дерев (=ансамблю) шляхом їх послідовного вирощування: кожне дерево росте, використовуючи інформацію з раніше вирощених дерев та відповідає зміненій версії вихідних даних. Ідея цієї процедури полягає в тому, що, на відміну від підгонки одного великого дерева до даних, що передбачає сильну підгонку даних і потенційне перенавчання, підхід boosting навчається повільно, а отже, ймовірність того, що наша модель побудує наступне дерево неправильно, менша. Алгоритм починається з побудови дерева регресії на вихідних даних і безперервно оновлює його відповідно до дерев регресії на залишках попередньої моделі. Враховуючи поточну модель, дерево встановлюється на залишках моделі, а не на змінній результату. Потім це нове дерево додається до підібраної функції, щоб оновити залишки. Під час boosting варто звернути увагу на те, що побудова кожного дерева сильно залежить від дерев, що вже були побудовані.

Алгоритм для побудови ансамблю регресійних дерев: boosting (Алгоритм 1.4.2)

a) Множина $\hat{f}(x) = 0$ і залишки $r_i = y_i$ для всіх i в множині навчальних даних
b) Для $b = 1; 2; \ldots; B$, повторюємо:
   1) Побудуйте дерево $\hat{f}^b$ з d розгалуженнями (d+1 вузли) на навчальних даних, використовуючи змінну r_i як залежну.
   2) Оновіть $\hat{f}$, додавши зменшену версію нового дерева: $\hat{f} \leftarrow \hat{f}(x) + \lambda \hat{f}^b(x)$
   3) Оновіть залишки: $r_i \leftarrow r_i + \lambda \hat{f}^b(x)$
c) Вивести boosted модель: $\hat{f}(x) = \sum_{b-1}^{B} \lambda \hat{f}^b(x)$

Алгоритм boosting вимагає специфікації трьох параметрів. По-перше, кількість кількість дерев - B. Boosting може призвести до перенавчання, якщо B завелике (хоча це перенавчання, як правило, відбувається повільно, якщо взагалі відбувається). По-друге, параметр shrink-age λ, невелике додатне

число. Він також відомий як швидкість навчання, оскільки він контролює величину, з якою кожне дерево робить внесок у модель, і зазвичай дорівнює 0,01 або 0,001. По-третє, кількість розгалужень у кожному дереві - d, що контролює складність або глибину взаємодії. [19]

Висновки за першим розділом

Було вивчено характеристику заробітної плати спеціалістів IT-сфери, які є економічними показниками. В результаті вивчення цих даних був зроблено висновок, що найбільш повне та точне прогнозування буде обчислено за допомогою такої науки, як економетрика, проводячи економічні дослідження. Були розглянуті економетричні методи прогнозування, , що дозволило обрати найбільш підходящий для даних метод. Враховуючи специфіку кожного представленого методу, ми можемо дійти до висновку, що нелінійна регресія є найбільш підходящою для пояснення даних та прогнозу. Для аналізу найбільш підходящим стане метод дисперсійного аналізу.

Розділ 2. Практична частина

З метою практичного вивчення побудови складної економетричної моделі та прогнозування була поставлена така задача: за допомогою прогнозування математичних моделей порівняти цей метод з реальними даними. Для виконання завдання будемо використовувати мову програмування R.

Для прогнозу за допомогою запропонованої форми функції заробітків нам потрібно виконати декілька кроків:

1. Розглянути тип та природу даних, підготувати їх для комфортної роботи.

2. Протестувати покроково моделі в залежності від наявності фіктивних змінних та їхньої взаємодії.

3. Запропонувати одну, що найбільш повно та якісно описує залежність заробітної плати від змінних.

4. Спрогнозувати дані на подальші півроку.

5. Порівняти з реальними даними.

2.1 Підготовка даних

Розпочнемо аналіз моделі з підготовки даних.

Вступ. Об’єктом дослідження було обрано зарплатню IT-спеціаліста, а саме software engineer, (за місяць у доларах після сплати податків).

Характеристика інформаційної бази для обґрунтування:

- Початкова база даних (за грудень 2021-го року):

https://github.com/devua/csv/blob/master/salaries/2021_june_raw.csv

Для прогнозування зарплатні ми відстежимо наступні фактори:

1. Стать
2. Освіта
3. Досвід
4. Локація
5. Англійська мова

Специфікація моделі:

Залежною змінною є SAL – заробітна плата IT-спеціаліста в доларах США за місяць.

Незалежні змінні:

1. SEX – стать спеціаліста.

2. AGE – вік спеціаліста (одиниця виміру: рік).
3. EDU – рівень освіти спеціаліста (одиниця виміру: рік).
4. EXP – загальний досвід роботи (одиниця виміру: рік).
5. ENG – рівень володіння англійською мовою спеціаліста.
6. LOC – місцезнаходження спеціаліста.

У нашій моделі присутні не тільки кількісні, але і якісні змінні.

Якісні, або категоріальні, змінні – це нечислові значення. Можемо зробити умовну класифікацію якісних змінних на три типи:

- Двійкові: змінні лише з двома категоріями.
- Номінальні: змінні, що можна організувати в більш ніж дві категорії, що не відповідають певному порядку.
- Порядкові: змінні, що можна організувати в більш ніж дві категорії, що відповідають певному порядку.

Відповідно, модель, що ми будуємо, має враховувати особливості роботи з цими змінними. Певну взаємодію факторів ми описали в теоретичній частині, однак їхня кількість може бути набагато більша. В силу того, що ми тестуємо різноманітні комбінації при спробі прийти до найбільш якісної моделі, нам необхідно розглянути усі можливі варіанти. Очевидно, перебір опцій вручну не є оптимальним, окрім того, існує ризик, що певна взаємодія не буде протестована.

Для розв’язання цієї проблеми використаймо розширення «xgbfi» в середовищі R для виявлення взаємодій функцій.

Використання коду розподілене на три кроки: початкова збірка моделі, експорт файлів, необхідних для xgbfi, і запуск xgbi.

Код запускає генерацію візуалізації дерева та файлу з описом перетинів факторів.

2.2 Побудова моделі покроково

Розпочнемо з аналізу запропонованого датасету нашим розширенням.
Взаємодія рівня 0/ Перший стовпчик „Interaction“ (тобто оцінка кожного з факторів окремо) запропонувала свої найбільш підходящі варіанти.

Interaction	Gain	FScore	wFScore	Average wFScore	Average Gain	Expected Gain	Gain Rank	FScore Rank	wFScore Rank	Avg wFScore Rank	Avg Gain Rank
f3	28134399558	20	5,90087073	0,295043537	1406719978	25283533272	1	1	1	1	1
f4	8143000188	15	2,406563965	0,160437598	542866679,2	1450676113	2	2	2	3	2
f9	977519040	2	0,096003572	0,048001786	488759520	14670046,44	3	6	6	6	3
f0	553727808	4	0,672471534	0,168117883	138431952	117888100,8	4	5	3	2	4
f1	279616064	9	0,435365037	0,048373893	31068451,56	14208783,76	5	3	4	5	6
f5	192619200	5	0,313239562	0,062647912	38523840	10815264,89	6	4	5	4	5

	Sex	Age	Education	Experience	English	Kyiv	Lviv	Odesa	Kharkiv	Dnipro
	f0	f1	f2	f3	f4	f5	f6	f7	f8	f9

(рис. 2.2.1)

Бачимо, що найпершим йде аргумент «Досвід», далі рівень англійської мови, окрім того, локація Дніпро, стать.

Розпочнемо побудову моделі з єдиної бінарної змінної, що є в цьому списку.

Визначимо спершу наявність або відсутність гендерної дискримінації. Введемо таку категорійну змінну:

$$female = \begin{cases} 1, \text{якщо індивідуум жінка} \\ 0, \text{якщо чоловік} \end{cases}$$

Ми включаємо цю змінну в модель оплати праці:

$$\ln Y_i = \beta_0 + \delta_0 \times female + u_i, \quad (2.2.1)$$

Якщо залежна зміна представлена у вигляді логарифму, то коефіцієнт фіктивної змінної, , помножений на 100, фіксує відсоткову різницю.

Якщо в цій моделі обчислити середню заробітну плату для чоловіків і жінок (не враховуючи освіту, досвід і т.і.), то маємо:

$$Females \rightarrow E(salary|female = 1,) = \beta_0 + \delta_0 \times 1$$

$$Males \rightarrow E(salary|female = 0,) = \beta_0 + \delta_0 \times 0$$

Звідси, $\delta_0 = E(salary|female = 1) - E(salary|male = 0)$

```
> sex_model <- lm( log(our_Data$Salary)  ~ our_Data$Sex)
> sum <- summary(sex_model)
> sum

Call:
lm(formula = log(our_Data$Salary) ~ our_Data$Sex)

Residuals:
    Min      1Q  Median      3Q     Max
-3.1803 -0.3770  0.1519  0.5086  4.1323

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.46966    0.03413 218.843   <2e-16 ***
our_Data$SexMale 0.31583    0.03584   8.812   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.697 on 4477 degrees of freedom
Multiple R-squared:  0.01705,   Adjusted R-squared:  0.01683
F-statistic: 77.65 on 1 and 4477 DF,  p-value: < 2.2e-16
```

(рис. 2.2.2)

Отже, оскільки відсоткова різниця приблизно дорівнює різниці логарифмів, помноженій на 100, це означає, що за оцінками моделі, жінки заробляють в середньому на 31,5% менше, ніж чоловіки.

Очевидно, що це не обов'язково відображає гендерну дискрімінацію як таку – з політичним підґрунтям, сексистськими настроями тощо. Ця різниця може бути поясена різноманітними факторами, що ми (поки що) не врахували у своїй моделі, а саме: освіта, досвід, вік тощо.

R^2 –коефіцієнт детермінації=0.01705. Бачимо, що сам по собі фактор дуже слабо пояснює модель, тобто є сенс розглянути його у взаємодії з більш значущим фактором.

Додамо до нашої моделі оплати праці змінну, що відповідає за функцією освіти:

$$\ln Y_i = \beta_0 + \delta_0 \times female + \beta_1 \times educ + u_i, \quad (2.2.2)$$

Якщо в цій моделі обчислити середню заробітну плату для чоловіків і жінок з урахуванням освіти, то маємо:

$$Females \rightarrow E(salary|female = 1, educ) = (\beta_0 + \delta_0 \times 1) + \beta_1 \times educ$$

$$Males \rightarrow E(salary|female = 0, educ) = \beta_0 + \delta_0 \times 0 + \beta_1 \times educ$$

```
> bi_model <- lm(log(our_Data$Salary) ~ our_Data$Sex + our_Data$fact_edu)
> summary(bi_model)

Call:
lm(formula = log(our_Data$Salary) ~ our_Data$Sex + our_Data$fact_edu)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2675 -0.3415  0.1337  0.4461  4.0561

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       7.24922    0.02680 270.522   <2e-16 ***
our_Data$Sex1    -0.32676    0.03410  -9.583   <2e-16 ***
our_Data$fact_edu  0.62345    0.02871  21.715   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.663 on 4476 degrees of freedom
Multiple R-squared:  0.1107,    Adjusted R-squared:  0.1103
F-statistic: 278.7 on 2 and 4476 DF,  p-value: < 2.2e-16

>
```

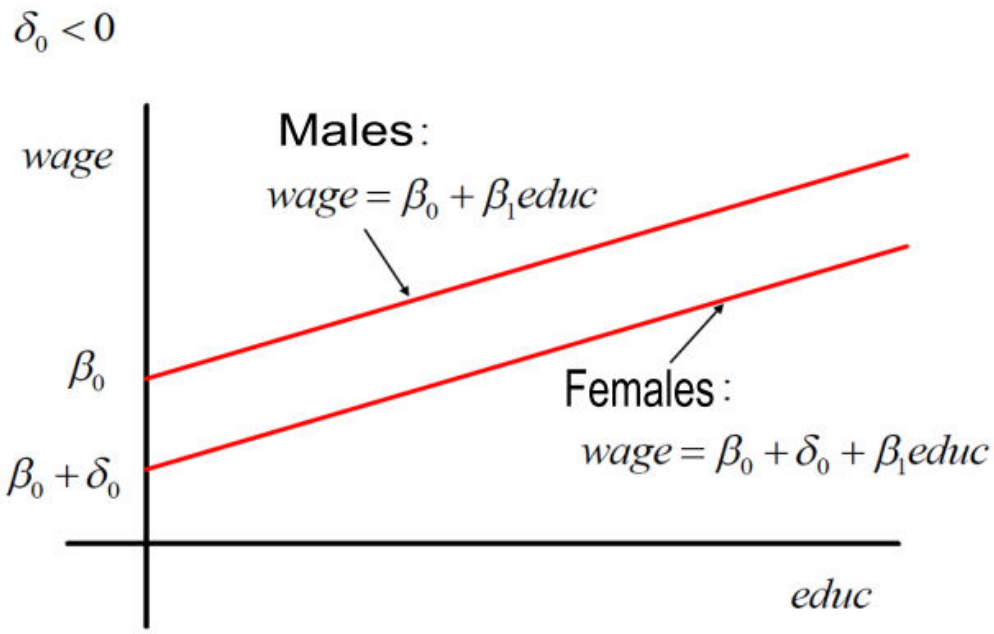
(рис. 2.2.3)

Звідси, $\delta_0 = E(salary|female = 1, educ) - E(salary|female = 0, educ)$

Отже, якщо

- $\delta_0 < 0$, то для одного й того самого рівня освіти середня ЗП жінок-програмісток нижча, аніж середня ЗП у чоловіків-програмістів.
- $\delta_0 = 0$, за однакового рівня освіти середня ЗП жінок не відрізняється від чоловіків.

Якщо $\delta_0 < 0$, то ми можемо графічно представити:



(рис. 2.2.4)

Надалі розглянемо взаємодію між двома бінарними змінними.

Змінна кількісті років освіти була перетворена у факторну наступним чином: люди з загальної кількості років освіти менше та 4 є категорією 0, більше - 1, тобто люди з закінченою вищою освітою та без.

$$\ln Y_i = \ln Y_0(\beta_0) + \delta_0 \times female + \beta_1 \times educ + \beta_3 (female \times edu)\, u_i \quad (2.2.3)$$

```
> # Yi = β0 + β1D1i + β2D2i + β3 (D1i × D2i ) + ui ;
> bi_model_1 <- lm(log(our_Data$Salary) ~  our_Data$Sex + our_Data$fact_edu + our_Data$Sex * our_Data$fact_edu)
> summary(bi_model_1)

Call:
lm(formula = log(our_Data$Salary) ~ our_Data$Sex + our_Data$fact_edu +
    our_Data$Sex * our_Data$fact_edu)

Residuals:
    Min      1Q  Median      3Q     Max
-3.2651 -0.3473  0.1361  0.4376  4.0332

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    7.26394    0.02781 261.187  < 2e-16 ***
our_Data$Sex1                 -0.50531    0.09689  -5.215 1.92e-07 ***
our_Data$fact_edu              0.60634    0.02999  20.220  < 2e-16 ***
our_Data$Sex1:our_Data$fact_edu  0.20377    0.10351   1.969   0.0491 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6628 on 4475 degrees of freedom
Multiple R-squared:  0.1115,    Adjusted R-squared:  0.1109
F-statistic: 187.2 on 3 and 4475 DF,  p-value: < 2.2e-16

> # друк підсумків коефіцієнтів
> coeftest(bi_model_1, vcov. = vcovHC, type = "HC1")

t test of coefficients:

                               Estimate Std. Error  t value  Pr(>|t|)
(Intercept)                    7.263935   0.031634 229.6208 < 2.2e-16 ***
our_Data$Sex1                 -0.505310   0.124544  -4.0573  5.05e-05 ***
our_Data$fact_edu              0.606344   0.033442  18.1312 < 2.2e-16 ***
our_Data$Sex1:our_Data$fact_edu  0.203774   0.130025   1.5672    0.1171
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

>
```

(рис. 2.2.5)

Бачимо, що p-value (рівень статистичної значущості) майже дорівнює 0.05, що не є найбільш потужним результатом. Тим не менш, така взаємодія покращила модель на соті, того можемо взяти її до уваги надалі.

Окремо розглянемо взаємодію **неперервної та двійкової змінної.**

$$\ln Y_i = \beta_0 + \beta_1 \times exp + \beta_2 \times educ + \beta_3 (exp \times edu) + u_i \quad (2.2.4)$$

Роки трудового стажу особи є неперервною змінною. Додаючи змінну взаємодії $exp \times edu$, ми дозволяємо впливу додаткового року досвіду роботи відрізнятися для осіб з вищою закінченою освітою та без неї.

```
> #a. Yi = β0 + β1D1i + β2D2i + ui ;
> bi_model_b1 <- lm(log(our_Data$Salary) ~ our_Data$Experience + our_Data$fact_edu +
+                     (our_Data$Experience * our_Data$fact_edu))
> summary(bi_model_b1)

Call:
lm(formula = log(our_Data$Salary) ~ our_Data$Experience + our_Data$fact_edu +
    (our_Data$Experience * our_Data$fact_edu))

Residuals:
    Min      1Q  Median      3Q     Max
-2.6369 -0.2728  0.0488  0.3420  4.1902

Coefficients:
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                               6.677196   0.028636 233.174   <2e-16 ***
our_Data$Experience                       0.208085   0.007674  27.114   <2e-16 ***
our_Data$fact_edu                         0.453991   0.032887  13.805   <2e-16 ***
our_Data$Experience:our_Data$fact_edu -0.067775   0.008154  -8.311   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5074 on 4475 degrees of freedom
Multiple R-squared:  0.4794,    Adjusted R-squared:  0.4791
F-statistic:  1374 on 3 and 4475 DF,  p-value: < 2.2e-16

> |
```

(рис. 2.2.6)

Коефіцієнт детермінації одразу значно виріс, щойно ми включили такий вирішальний фактор, як досвід. Абсолютно усі зміні є статистично значущими, їхня взаємодія має так само вагомий вплив.

Враховуючи, що ми не обмежуємося однією лишень логарифмічно-лінійною моделлю, можемо розглянути модель log-log.

$$\ln Y_i = \beta_0 + \ln(\beta_1 exp) + \ln(\beta_2 (exp \times edu)) + u_i \quad (2.2.4)$$

Ця модель стверджує, що очікуваний вплив додаткового року досвіду роботи на заробіток відрізняється для випускників університетів і тих, хто ще не закінчили навчання, але самостійне навчання не збільшує прибутки.

```
> #e. ln(Yi ) = β0 + β1 ln(Xi ) + β2 (ln(Xi ) × Di ) + ui
> bi_model_e <- lm(log(our_Data$Salary) ~  log(our_Data$Experience) + log(our_Data$Experience) * our_Data$fact_edu)
> summary(bi_model_e)

Call:
lm(formula = log(our_Data$Salary) ~ log(our_Data$Experience) +
    log(our_Data$Experience) * our_Data$fact_edu)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6311 -0.2240  0.0374  0.2693  4.1407

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                              7.00273    0.01825 383.681  < 2e-16 ***
log(our_Data$Experience)                 0.54540    0.01477  36.939  < 2e-16 ***
our_Data$fact_edu                        0.06524    0.02235   2.920  0.00352 **
log(our_Data$Experience):our_Data$fact_edu  0.02200    0.01678   1.311  0.19002
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4293 on 4475 degrees of freedom
Multiple R-squared:  0.6272,    Adjusted R-squared:  0.627
F-statistic:  2510 on 3 and 4475 DF,  p-value: < 2.2e-16

> |
```

(рис. 2.2.7)

Бачимо, що поки що ця модель є найбільш успішною з точки зору рівня пояснення залежної зміни незалежними.

Наша мета: наблизится до запропонованої модифікованої моделі та перевірити, чи дійсно вона є кращою за інші.

$$\ln Y_i = \beta_0 + \delta_2 R_2 + \cdots + \delta_j R_j + \alpha_1 C_{1i} + \alpha_G D_{Gi} + \beta_1 s_i + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 s_i X_i + u_i$$

Подальші інтеракції перевіряємо за допомогою «xgbi».

Взаємодія рівня 1 пропонує використати $(exp \times exp) = X_i^2$ (як і було розглянуто в теоретичній частині), тобто досвід в квадраті. Окрім того, варто перевірити взаємодію досвіду і рівня англійської мови. Також певним чином впливає комбінація статі з рівнем англійської. Вплив Дніпра в дуеті зі статтю та англійським має місце бути окремим фактором.

Interaction	Gain	FScore	wFScore	Average wFScore	Average Gain	Expected Gain	Gain Rank	FScore Rank	wFScore Rank
f3|f3	51661923560	8	2,553471757	0,31918397	6457740445	25527642535	1	2	2
f3|f4	14181111556	16	2,614646126	0,163415383	886319472,3	4713741671	2	1	1
f4|f9	5454939072	2	0,08506363	0,042531815	2727469536	15531160,26	3	8	11
f0|f4	3101477308	4	0,659745479	0,16493637	775369327	788800618,4	4	4	3
f4|f4	2946161660	2	0,403661532	0,201830766	1473080830	594626064	5	9	5
f0|f9	1118354240	1	0,012726055	0,012726055	1118354240	14232237,48	6	12	13
f0|f3	1067153280	3	0,592319714	0,197439905	355717760	186752633,1	7	6	4
f4|f5	498387072	2	0,097343157	0,048671578	249193536	24257285,49	8	10	9
f1|f3	431929280	4	0,250948873	0,062737218	107982320	25636692,46	9	5	7
f3|f5	351439312	5	0,283768698	0,05675374	70287862,4	23124047,62	10	3	6
f1|f4	226650576	3	0,150480018	0,050160006	75550192	20495483,56	11	7	8
f1|f1	145577984	1	0,031926769	0,031926769	145577984	4647834,72	12	13	12
f1|f5	22630512	2	0,088412592	0,044206296	11315256	1000411,113	13	11	10

Sex	Age	Education	Experience	English	Kyiv	Lviv	Odesa	Kharkiv	Dnipro
f0	f1	f2	f3	f4	f5	f6	f7	f8	f9

(рис. 2.2.8)

Отже, перевіримо вищезапропоновані інтеракції в RStudio.

$$\ln Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + u_i = \ln Y_0 + \beta_1 exp + \beta_2 exp^2 + u_i \quad (2.2.5)$$

Коефіцієнт детермінації тут майже такий самий, як і у моделі з освітою та включеною взаємодією освіти та досвіду.

```
> bi_model <- lm(log(our_Data$Salary) ~ our_Data$Experience + our_Data$Experience*our_Data$Experience
> summary(bi_model)

Call:
lm(formula = log(our_Data$Salary) ~ our_Data$Experience + our_Data$Experience *
    our_Data$Experience)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6506 -0.2856  0.0535  0.3616  4.2716

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)         7.018082   0.014308  490.49   <2e-16 ***
our_Data$Experience 0.156158   0.002546   61.34   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5182 on 4477 degrees of freedom
Multiple R-squared:  0.4566,    Adjusted R-squared:  0.4565
F-statistic:  3762 on 1 and 4477 DF,  p-value: < 2.2e-16
```

(рис. 2.2.9)

$$\ln Y_i = \beta_0 + \beta_1 exp + \beta_2 eng + \beta_3 (eng \times \exp) + u_i \quad (2.2.6)$$

```
> bi_model <- lm(log(our_Data$Salary) ~ our_Data$Experience + our_Data$English + our_Data$English*our_Data$Experienc
e)
> summary(bi_model)

Call:
lm(formula = log(our_Data$Salary) ~ our_Data$Experience + our_Data$English +
    our_Data$English * our_Data$Experience)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6376 -0.2935  0.0513  0.3437  4.1403

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                         6.591288   0.038602 170.751   <2e-16 ***
our_Data$Experience                 0.158213   0.007089  22.320   <2e-16 ***
our_Data$English                    0.192839   0.015815  12.193   <2e-16 ***
our_Data$Experience:our_Data$English -0.004086   0.002765  -1.478     0.14
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.497 on 4475 degrees of freedom
Multiple R-squared:  0.5005,    Adjusted R-squared:  0.5002
F-statistic:  1495 on 3 and 4475 DF,  p-value: < 2.2e-16
```

(рис. 2.2.10)

Нова зміна, що відповідає за спільний вплив англійської та досвіду виявилася статистично незначущою, тож на майбутнє не будемо використовувати цей фактор. Натомість перевіримо комбінацію англійської та статі.

$$\ln Y_i = \beta_0 + \beta_1 sex + \beta_2 eng + \beta_3 (eng \times sex) + u_i \quad (2.2.7)$$

```
> bi_model <- lm(log(our_Data$Salary) ~ our_Data$Sex + our_Data$English + our_Data$English*our_Data$Sex)
> summary(bi_model)

Call:
lm(formula = log(our_Data$Salary) ~ our_Data$Sex + our_Data$English +
    our_Data$English * our_Data$Sex)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3436 -0.3479  0.1451  0.4631  4.0448

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                    7.13932    0.03057 233.506   <2e-16 ***
our_Data$Sex                  -0.22633    0.11260  -2.010   0.0445 *
our_Data$English               0.26981    0.01201  22.458   <2e-16 ***
our_Data$Sex:our_Data$English -0.05507    0.04168  -1.322   0.1864
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.659 on 4475 degrees of freedom
Multiple R-squared:  0.1217,    Adjusted R-squared:  0.1211
F-statistic: 206.7 on 3 and 4475 DF,  p-value: < 2.2e-16
```

(рис. 2.2.11)

Коефіцієнт детермінації доволі низький, так само, як і p-value нової змінної інтеракції. Отже, можемо не враховувати цю опцію.

Фіктивні зміни з кількома категоріями

Найбільш слизький момент нашої моделі – місцезнаходження айтівців. Ми хочемо довести або спростувати гіпотезу, що локація програміста має безпосередній вплив на зарплатню. Для аналізу обрали вибірку з 5 найбільших міст України, без прив’язки до статі, досвіду, враховуючи лише те, що у них однакова освіта.

Розглянемо наступну модель:

$$\ln Y_i = \beta_0 + \delta_2 \times Lviv + \delta_3 \times Dnipro + \delta_3 \times Kharkiv + \delta_3 \times Odesa + \beta_1 \times edu + u_i, \quad (2.2.8)$$

де категорією, від якої ми відштовхуємося, є Київ.

Послуговуючись теорією, була використана запропонована модель. Оцінив, ми отримаємо наступний результат:

```
> bi_model <- lm(log(our_Data$Salary) ~ our_Data$Lviv + our_Data$Dnipro + our_Data$Odesa + our_Data$Kharkiv )
> summary(bi_model)

Call:
lm(formula = log(our_Data$Salary) ~ our_Data$Lviv + our_Data$Dnipro +
    our_Data$Odesa + our_Data$Kharkiv)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0945 -0.3588  0.1521  0.4983  4.0082

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      7.85430    0.01439 545.910  < 2e-16 ***
our_Data$Lviv   -0.18603    0.02803  -6.638 3.56e-11 ***
our_Data$Dnipro -0.26055    0.04047  -6.438 1.33e-10 ***
our_Data$Odesa  -0.15465    0.04470  -3.460 0.000545 ***
our_Data$Kharkiv -0.22108    0.02998  -7.374 1.96e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6954 on 4474 degrees of freedom
Multiple R-squared:  0.02228,   Adjusted R-squared:  0.02141
F-statistic: 25.49 on 4 and 4474 DF,  p-value: < 2.2e-16
```

(рис. 2.2.12)

Бачимо, що Дніпро дійсно має найбільшу різницю в порівнянні з Києвом, а отже, ми можемо лишити це місцезнаходження як важливе.

Взаємодія рівня 2 пропонує перелік різноманітних комбінацій, з яких $sex \times exp \times eng$ має найбільший сенс.

Interaction	Gain	FScore	wFScore	Average wFScore	Average Gain	Expected Gain	Gain Rank	FScore Rank	wFScore Rank	Avg wFScore Rank
f3|f3|f4	66394524528	10	2,032373298	0,20323733	6639452453	21443380949	1	1	1	1
f3|f3|f3	42314304230	4	0,553471757	0,138367939	10578576058	6023388103	2	4	4	3
f3|f4|f4	11874712956	8	0,98593436	0,123241795	1484339120	2031122392	3	2	3	4
f0|f3|f4	10828458940	6	1,152266131	0,192044355	1804743157	2700089321	4	3	2	2
f0|f4|f9	5556357440	1	0,001786113	0,001786113	5556357440	9924282,099	5	10	16	16
f0|f3|f3	4667897726	1	0,099799062	0,099799062	4667897726	465851815,9	6	11	8	5
f4|f4|f5	3038138492	2	0,097343157	0,048671578	1519069246	147870996	7	7	9	11
f1|f3|f4	2108493504	4	0,279973208	0,069993302	527123376	178980125,5	8	5	5	7
f3|f3|f5	1626865616	4	0,224157178	0,056039294	406716404	120356137,7	9	6	6	9
f0|f3|f9	1465124672	1	0,012726055	0,012726055	1465124672	18645257,04	10	12	15	15
f1|f3|f3	1222718656	2	0,105827194	0,052913597	611359328	64698441,95	11	8	7	10
f3|f4|f5	696045568	1	0,05961152	0,05961152	696045568	41492334,6	12	13	12	8
f1|f1|f4	280787968	1	0,031926769	0,031926769	280787968	8964652,696	13	14	13	13
f3|f4|f9	234132480	1	0,083277517	0,083277517	234132480	19497971,65	14	15	11	6
f0|f1|f4	180332432	1	0,015628489	0,015628489	180332432	2818323,34	15	16	14	14
f1|f3|f5	153104496	2	0,088412592	0,044206296	76552248	6768182,676	16	9	10	12

Sex	Age	Education	Experience	English	Kyiv	Lviv	Odesa	Kharkiv	Dnipro
f0	f1	f2	f3	f4	f5	f6	f7	f8	f9

(рис. 2.2.13)

Можемо додати цю інтеракцію до моделі (2.2.4), яка поки що виявилася найуспішнішою з усіх вище перерахованих.

$$\ln Y_i = \beta_0 + \ln(\beta_1 exp) + \ln(\beta_2 (exp \times edu)) + \beta_3 (sex \times exp \times eng) + u_i \quad (2.2.9)$$

Дійсно, ця взаємодія є вирішальною для даного випадку.

```
> bi_model <- lm(log(our_Data$Salary) ~ log(our_Data$Experience) + our_Data$fact_edu*our_Data$Experience + our_Data
$Experience*our_Data$fact_edu*our_Data$Sex)
> summary(bi_model)

Call:
lm(formula = log(our_Data$Salary) ~ log(our_Data$Experience) +
    our_Data$fact_edu * our_Data$Experience + our_Data$Experience *
    our_Data$fact_edu * our_Data$Sex)

Residuals:
    Min      1Q  Median      3Q     Max
-2.6298 -0.2249  0.0475  0.2706  4.1994

Coefficients:
                                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                                       7.034760   0.026541 265.051   <2e-16 ***
log(our_Data$Experience)                          0.594656   0.014212  41.841   <2e-16 ***
our_Data$fact_edu                                 0.066652   0.030580   2.180   0.0293 *
our_Data$Experience                              -0.016675   0.008364  -1.994   0.0463 *
our_Data$Sex                                     -0.223529   0.091773  -2.436   0.0149 *
our_Data$fact_edu:our_Data$Experience             0.005254   0.007198   0.730   0.4655
our_Data$Experience:our_Data$Sex                  0.097896   0.051149   1.914   0.0557 .
our_Data$fact_edu:our_Data$Sex                    0.182448   0.101322   1.801   0.0718 .
our_Data$fact_edu:our_Data$Experience:our_Data$Sex -0.121448   0.051854  -2.342   0.0192 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4271 on 4470 degrees of freedom
Multiple R-squared:  0.6315,    Adjusted R-squared:  0.6309
F-statistic: 957.7 on 8 and 4470 DF,  p-value: < 2.2e-16
```

(рис. 2.2.13)

Сама по собі інтеракція освіти та досвіду підвищує рівень зарплати, проте комбінація жінка+освіта+досвід негативно відображається на моделі – рівень ЗП зменшується на 12.1%.

Отже, проаналізувавши кожний фактор окремо, попарні та потрійні комбінації, ми можемо переходити до побудови «ідеальної» моделі.

Модель (1.3.10), що була запропонована в теоретичній частині як найбільш повна, може бути розширена до

$$\ln Y_i = \beta_0 + \delta_2 R + \alpha_G sex + \beta_1 edu + \ln(\beta_2 exp) + \beta_3 exp^2 + \ln(\beta_4 edu \times \exp) + \beta_5 eng + \beta_6\,(sex\ \times exp\ \times eng) +\ u_i \quad (2.2.10)$$

Модель (2.2.10) і є новою модифікацією, що, за попередніми припущеннями, має бути більш успішною за попередні.

Для неї і зробимо повний статистичний аналіз.

2.3 Широкий статистичний аналіз запропонованої моделі

Розглянемо модель (2.2.10) та спершу поглянемо на загальний аналіз.

```
> View(our_Data)
> bi_model_coef <- lm(log(our_Data$Salary) ~ our_Data$Dnipro + our_Data$Sex + our_Data$Education+ log(our_Data$Expe
rience) +  our_Data$EXP_SQR + our_Data$Experience*our_Data$Education +our_Data$English + (our_Data$Sex*our_Data$Exp
erience*our_Data$Education) )
> summary(bi_model_coef)

Call:
lm(formula = log(our_Data$Salary) ~ our_Data$Dnipro + our_Data$Sex +
    our_Data$Education + log(our_Data$Experience) + our_Data$EXP_SQR +
    our_Data$Experience * our_Data$Education + our_Data$English +
    (our_Data$Sex * our_Data$Experience * our_Data$Education))

Residuals:
    Min      1Q  Median      3Q     Max
-2.5393 -0.2014  0.0257  0.2347  4.3002

Coefficients:
                                                    Estimate Std. Error t value Pr(>|t|)
(Intercept)                                        6.3341696  0.0427721 148.091  < 2e-16 ***
our_Data$Dnipro                                   -0.0928815  0.0223535  -4.155 3.31e-05 ***
our_Data$Sex                                      -0.3769274  0.1274968  -2.956 0.003129 **
our_Data$Education                                 0.0096642  0.0062013   1.558 0.119206
log(our_Data$Experience)                           0.3119969  0.0211236  14.770  < 2e-16 ***
our_Data$EXP_SQR                                  -0.0186176  0.0011733 -15.867  < 2e-16 ***
our_Data$Experience                                0.2657229  0.0187280  14.189  < 2e-16 ***
our_Data$English                                   0.1503550  0.0070424  21.350  < 2e-16 ***
our_Data$Education:our_Data$Experience            -0.0008783  0.0011833  -0.742 0.457985
our_Data$Sex:our_Data$Experience                   0.1220587  0.0433790   2.814 0.004918 **
our_Data$Sex:our_Data$Education                    0.0518323  0.0225951   2.294 0.021839 *
our_Data$Sex:our_Data$Education:our_Data$Experience -0.0241220  0.0070976  -3.399 0.000683 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3943 on 4467 degrees of freedom
Multiple R-squared:  0.6861,    Adjusted R-squared:  0.6854
F-statistic: 887.7 on 11 and 4467 DF,  p-value: < 2.2e-16
```

(рис. 2.3.1)

Наша модель доволі вдало пояснюється цими факторами – **коефіцієнт детермінації становить 0.68**, що означає, що модель можна описати на 68% за допомогою цих змінних.

Подивимося на p-value кожного з атрибутів. Наш рівень значущості $p<0.05$.

Саме з цього аналізу дуже добре прослідковується важливість глибоких взаємодій. Інтеракції «Стать-Освіта» є статистично незначущою, адже p-значення є більшим за 0.05. Тим не менш, взаємодія «Стать-Освіта-Досвід» має вельми низьке p-значення і є однією з найбільших значущих в цій моделі.

Сформулюємо гіпотези *H*0 та *H*1. Пропонуються наступні гіпотези:

- Нульова H0 - усі коефіцієнти β_i дорівнюють 0.
- Альтернативна H1 - коефіцієнти дорівнюють числу, відмінному від нуля, а отже, самі показники є статистично значущими.

Знаходимо **наше критичне значення** (|t_kr|)за допомогою таблиці (число ступенів свободи визначаємо з summary)

1) t-статистика всіх коефіцієнтів для рівня значущості 5%;

Число степеней свободы $f = n - 1$	n	Доверительная вероятность			
		0.90	0.95	0.99	0.999
1	2	6.3137515148	12.7062047364	63.6567411629	636.619249432
2	3	2.91998558036	4.30265272991	9.92484320092	31.599054577
3	4	2.3533634348	3.18244630528	5.84090929976	12.9239786366
4	5	2.13184678134	2.7764451052	4.60409487142	8.61030158138
5	6	2.01504837267	2.57058183661	4.03214298356	6.86882663987
6	7	1.94318028039	2.44691184879	3.70742802132	5.95881617993
7	8	1.89457860506	2.36462425101	3.49948329735	5.40788252098
8	9	1.85954803752	2.30600413503	3.35538733133	5.04130543339
9	10	1.83311293265	2.26215716274	3.24983554402	4.78091258593
10	11	1.81246112281	2.22813885196	3.16927266718	4.5868938587
11	12	1.7958848187	2.20098516008	3.10580651322	4.43697933823
12	13	1.78228755565	2.17881282966	3.05453958834	4.31779128361
13	14	1.77093339599	2.16036865646	3.01227583821	4.22083172771
14	15	1.76131013577	2.14478668792	2.97684273411	4.14045411274
15	16	1.75305035569	2.13144954556	2.94671288334	4.0727651959
16	17	1.74588367628	2.11990529922	2.92078162235	4.0149963326
17	18	1.73960672608	2.10981557783	2.89823051963	3.96512626361
18	19	1.73406360662	2.10092204024	2.87844047271	3.92164582001
19	20	1.72913281152	2.09302405441	2.86093460645	3.88340584948
20	21	1.72471824292	2.08596344727	2.84533970978	3.84951627298
21	22	1.72074290281	2.07961384473	2.83135955802	3.81927716303
22	23	1.71714437438	2.0738730679	2.8187560606	3.79213067089
23	24	1.71387152775	2.06865761042	2.80733568377	3.76762680377
24	25	1.71088207991	2.06389856163	2.79693950477	3.74539861893
25	26	1.70814076125	2.05953855275	2.78743581368	3.72514394948
26	27	1.70561791976	2.05552943864	2.77871453333	3.70661174331
27	28	1.70328844572	2.05183051648	2.77068295712	3.68959171334
28	29	1.70113093427	2.0484071418	2.76326245546	3.67390640062
29	30	1.69912702653	2.04522964213	2.75638590367	3.6594050194
30	31	1.69726089436	2.0422724563	2.74999565357	3.645958635
40	41	1.68385101139	2.021075383	2.70445926743	3.55096576086
60	61	1.67064886465	2.00029782106	2.66028303115	3.4602004692
120	121	1.6576508993	1.97993040505	2.61742114477	3.37345376507
999999.0	1000000.0	1.6448551507	1.95996635682	2.57583422011	3.29053646126

(рис. 2.3.2)

Орієнтуючись на 4467 степенів свободи, ми знаходимо t-критичне=1.96.

Оскільки наші значення t усюди є більшими за |t_kr|, окрім змінних безпосередньо освіти та взаємодії освіти та досвіду, то робимо висновок, що коефіцієнти відрізняються від нуля. Відповідно, приймаємо альтернативну гіпотезу H1 (β_i !=0) для усіх інших коефіцієнтів, для вищезазначених запевнюємося в тому, що вони дійсно дуже близькі до нуля.
Така різка зміна значущості в порівнянні з іншими моделями, розглянутими раніше, пояснена тим, що в запропонованій моделі з’явилися більш сильні значимі предиктори. Зміні взаємодії, окрім всього, включають в себе фактор «досвід», який є одним з найбільш вагомих у моделі.

Розглянемо довірчі інтервали для коефіцієнтів моделі:

```
> #довірчі інтервали для коефіцієнтів моделі;
> confint(bi_model_coef)
                                                                2.5 %       97.5 %
(Intercept)                                                6.250315164  6.418024074
our_Data$Dnipro                                           -0.136705491 -0.049057437
our_Data$Sex                                              -0.626884177 -0.126970596
our_Data$Education                                        -0.002493425  0.021821835
log(our_Data$Experience)                                   0.270584253  0.353409572
our_Data$EXP_SQR                                          -0.020917890 -0.016317319
our_Data$Experience                                        0.229006687  0.302439055
our_Data$English                                           0.136548506  0.164161547
our_Data$Education:our_Data$Experience                    -0.003198070  0.001441537
our_Data$Sex:our_Data$Experience                           0.037014306  0.207103099
our_Data$Sex:our_Data$Education                            0.007534669  0.096129949
our_Data$Sex:our_Data$Education:our_Data$Experience       -0.038036727 -0.010207182
```

(рис. 2.3.3)

2.4 Прогноз на наступний період

Для перевірки запропонованої модифікації моделі ми візьмемо аналогічний датасет за наступний період. Дані для основного набору збиралися сайтом DOU за травень 2021 року, тренувальний датасет був взятий звідси ж, тільки за грудень 2021.

Завдяки цим цифрам ми можемо оцінити, наскільки вдалим та точним є наш прогноз.

Отже, були отримані такі результати:

(Табл. 2.4.1)

SALARY PREDICTED	SALARY REAL
4975,20	18000
5087,05	15000
4697,32	14000
5469,07	13500
4972,13	13000
4156,97	13000
5799,41	11700
5459,44	11000
2770,52	10800
4652,15	10600
4697,32	10500
4697,32	10400
4020,26	10000
4697,32	10000
4041,58	10000
5459,44	10000
4156,97	9700
5406,94	9500
4697,32	9500
5087,05	9500
4697,32	9100
498,98	9000
4697,32	9000
5406,94	9000
4975,20	9000
4972,13	9000
4705,61	9000
4697,32	9000
4697,32	9000
5459,44	9000
4697,32	9000
4697,32	8700
4697,32	8670
5406,94	8600
4697,32	8600

Бачимо, що на перших 40 кейсах дані доволі сильно розходяться, окрім того, є аномальні значення.

Проте при порівнянні реальної ЗП менше за 5000, дані стають доволі схожими і близькими до правди.

Прогнози збігаються реальними даними, допоки зарплатня не стає менша за 1000.

(Табл. 2.4.2)

4697,32	4700
4697,32	4700
4697,32	4655
4697,32	4655
3040,18	4650
4002,72	4650
4697,32	4605
3533,44	4600
3240,19	4600
3576,66	4600
2963,18	4600
4418,11	4600
4156,97	4600
4201,04	4600
4972,13	4600
2549,53	4600
4652,15	4600
4652,15	4600
4972,13	4600
5912,41	4600
4091,73	4600
3683,10	4600
4693,22	4600
4293,26	4600
4697,32	4600
4972,13	4600
4972,13	4600
5087,05	4600
4697,32	4600
4705,61	4600
4041,58	4600
5087,05	4600
4041,58	4587
5406,94	4575
2440,08	4560
4002,72	4560

(Табл. 2.4.3)

1891,25	1200
1818,77	1200
1806,37	1200
1709,28	1185
2440,08	1184
1193,99	1180
1374,21	1150
1387,71	1150
1125,03	1140
1193,99	1140
1776,13	1132
1585,93	1100
995,14	1100
1162,93	1100
1412,31	1100
1193,99	1100
1351,61	1100
1709,28	1100
1243,51	1100
1445,26	1100
1387,71	1100
1627,23	1100
1507,96	1100
1116,33	1100
1116,33	1100
1017,32	1100
1346,42	1100
2440,08	1100
283,60	1100
1346,42	1100
1270,04	1100
1323,02	1100
1162,93	1100
2440,08	1100
2113,86	1100
2036,98	1100

Щойно справжня ЗП стає менша за 1000, наша модель починає себе вести неадекватно та давати дані, що зовсім не відповідають дійсності.

(Табл. 2.4.4)

4156,97	380
1340,22	380
6036,82	370
7925,55	350
7537,08	350
6484,92	350
6303,76	350
1088,73	350
7326,52	350
7952,27	345
1070,60	335
8328,53	330
995,14	330
7016,28	330
7326,52	325
455,75	320
455,75	300
6842,16	300
337,39	300
4153,28	300
8609,08	300
3450,60	300
7326,52	300
1579,91	300
7605,11	300
7326,52	300
8562,24	300
3472,24	300
5091,96	300
8562,24	300
4666,62	300
3472,24	300
101,37	300
1193,99	300
1284,03	300
4690,36	300
4015,17	300
283,60	300
455,75	300
4690,36	300
455,75	280
4381,14	250
5453,72	250

Звідси робимо висновок, що модель працює для середнього рівня зарплат, що обумовлюється пересічним рівнем досвіду та освіти.

Для дійсно високих і низьких зарплат модель не видає очікуваних результатів. Отже, запропонована модифікація потребує доопрацювання з урахуванням низки особливостей, що і спричинюють такі результати.

Ймовірніше всього, це є виняткові персональні характеристики, що зазвичай не збираються в таких опитувальниках, проте часто прямо корелюють з рівнем зарплат. Наприклад, працелюбність, кмітливість, гострий розум, нестандартне мислення, відданість роботі і т.і. Ці специфічні здібності зазвичай є тою нерозв’язаною задачею, над якою б’ються дослідники протягом століття. Окрім них, є ще багато факторів, які мають вплив саме на зарплати в IT-сфері, що їх варто розглядати детально в контексті.

Висновки

У результаті розбору теорії та її застосуванні на практиці було ґрунтовно досліджено нелінійну регресійну модель, методи її опрацювання та прогнозування.

У першому розділі кваліфікаційної роботи було розглянуто термінологію та основні поняття та методи, необхідні для подальшого вивчення та прогнозування нелінійної регресійної множинної моделі. Особливу увагу в роботі було приділено log-linear і log-log моделям.

У другому розділі був розроблений алгоритм для реалізації аналізу та прогнозування, запрограмований на мові R. Була наведена запропонована модифікація моделі, проведено її аналіз та спрогнозовано дані на наступний період.

Отримано наступні результати:

- визначено фактори, що формують модель заробітної плати і є статистично значущими;
- досліджено взаємодію між бінарними, категоріальними та ординальними змінними. Проведено аналіз взаємодій факторів та обрано найбільш вагомі;
- опрацьовано алгоритм boosting для побудови ансамблю регресійних дерев
- розроблена модель прогнозу та проведено порівняння спрогнозованих значень з реальними за наступний період.

Список використаної літератури

1. Ринок праці під час війни. Ірина Іполітова
2. Калина А.В., Конева М.И., Ященко В.А. Современный экономический анализ и прогнозирование: учебно-методическое пособие. – К. : Межрегиональная Академия управления персоналом, 1997. – 272 с.
3. Бутакова М.М. . Економічне прогнозування: методи і прийоми практичних розрахунків: навчальний посібник / М.М. Бутакова. - 2-е вид., Испр. - М.: КНОРУС. - 168 с., 2010
4. Ферстер Е., Ренц Б. Методи кореляційного та регресійного аналізу .- М.: Фінанси і статистика, 1983
5. Єлісєєва І.І.. -М: Фінанси і статистика, 2001
6. Д. Хендрі. Econometrika: alхімія чи наука? (рус.) // Ековест. - 2003. - № 2. - С. 172-196.
7. Бернд Э. Практика эконометрики: классика и современность. — М.: Юнити-Дана, 2005. — 848 с
8. Норкін Богдан Володимирович. Метод послідовних наближень для розв'язання інтегральних рівнянь актуарної математики : дис... канд. фіз.-мат. наук: 01.05.01 / НАН України; Інститут кібернетики ім. В.М.Глушкова. - К., 2006.
9. Statistical Models: Theory and Practice. David A. Freedman (2009).
10. Моргенштерн О. О точности экономико-статистических наблюдений. — М.: Статистика, 1968. — 324 с
11. Box, George; Jenkins, Gwilym (1976). Time Series Analysis: forecasting and control, rev. ed. Oakland, California: Holden-Day
12. Hamilton, James (1994). Time Series Analysis. Princeton University Press. ISBN 0-691-04289-6.
13. Weigend A. S., Gershenfeld N. A. (Eds.) (1994), Time Series Prediction: Forecasting the Future and Understanding the Past. Proceedings of the NATO Advanced Research Workshop on Comparative Time Series Analysis (Santa Fe, May 1992), Addison-Wesley. (англ.)
14. Linear Regression Models with Logarithmic Transformations. Kenneth Benoit∗. Methodology Institute London School of Economics kbenoit@lse.ac.uk March 17, 2011.
15. Gelman, Andrew (2005). "Analysis of variance? Why it is more important than ever". The Annals of Statistics. 33:
16. Statistical Methods of Analysis. 3a Chin Long Chiang
17. Statistics and Introduction to Econometrics M. Angeles Carnero Departamento de Fundamentos del Análisis Económico Year 2014-15

18. Applied Econometrics Labor Econometrics Michael Ash
19. Feature Interactions in XGBoost Kshitij Goyal, Sebastijan Dumancic, Hendrik Blockeel KU Leuven, Belgium

Додаток до 2.2

```
install.packages("xgboost")
install.packages("Ecdat")
install.packages("caret")
install.packages('DiagrammeR')
install.packages("DiagrammeRsvg")
install.packages("rsvg")
install.packages("dplyr")
install.packages("lmtest")
install.packages("sandwich")


library(GGally)
library(ggplot2)
library(ggpairs)
library(car)
library(sandwich)
library(lmtest)
library(RFGLS)
library(nlme)
library(rr2)
library(sandwich)
library(lmtest)
library(DiagrammeR)
library(xgboost)
library(Ecdat)
library(caret)
library(DiagrammeRsvg)
library( rsvg)
library(dplyr)



library(readxl)
our_Data <- read_excel("our_Data.xlsx")
str(our_Data)
```

```
Encoding(colnames(our_Data)) <- "UTF-8"


our_Data$Location <- NULL

our_Data$Position <- NULL

our_Data$Specialisation <- NULL

our_Data$Role <- NULL

our_Data$`Main programming language` <- NULL

our_Data$Uni <- NULL


str(our_Data)


plot(our_Data$Age, our_Data$Salary,

pch = 20,

col=2,

main="")


#_____________________________________________________________________-


df <- data.matrix(our_Data, rownames.force = NA)

nrow(df)


sal <- df[, 6]

sal


df_x <- df[,colnames(df)!="Salary"]

df_x


bst <- xgboost(data = df_x, label = sal, max.depth = 5, eta = 3, nthread = 2,
nround = 2, objective = "reg:linear")


pdf(file="C:/Olha/Applied Maths 4/Bachelor thesis/My plot.pdf", width=5,
height=5)

xgb.plot.tree(feature_names = names((df_x)), model = bst)

dev.off()


help(xgboost)
```

```
tree_plot <- xgb.plot.tree(feature_names = names((df_x)), model = bst)
# export plot object to file
export_graph(tree_plot, "xgboost_tree_plot.pdf", width = 1000, height = 1000)



xgb.importance(colnames(df_x), model = bst)

featureList <- names(df_x)
featureVector <- c()
for (i in 1:length(featureList)) {
  featureVector[i] <- paste(i, featureList[i], "q", sep="\t")
}
write.table(featureVector, "fmap.txt", row.names=FALSE, quote = FALSE,
col.names = FALSE)
xgb.dump(model = bst, fname = 'xgb.dump', fmap = "fmap.txt", with_stats =
TRUE)
#Візьмемо дві бінарні змінні D1i  і D2i
#D1i - 1, якщо особа має вищу освіту, 0 якщо не має
#D2i - 1, якщо і жінка, 0, якщо чоловік
#Yi = ln(salary)

#додамо змінну "факторна освіта"
our_Datafact_edu <- as.numeric(our_DataEdu >= 2)

our_DataExperience <- as.numeric(our_DataExperience)
str(our_DataExperience)
our_DataSex

bi_model <- lm(our_DataSalary ~ our_DataSex + our_Datafact_edu)
summary(bi_model)
# друк підсумків коефіцієнтів
coeftest(bi_model, vcov. = vcovHC, type = "HC1")
```

```
# Yi = β0 + β1D1i + β2D2i + β3 (D1i × D2i ) + ui ;


bi_model_1 <- lm(our_DataSalary ~  our_DataSex + our_Datafact_edu +
our_DataSex * our_Datafact_edu)


summary(bi_model_1)

# друк підсумків коефіцієнтів

coeftest(bi_model_1, vcov. = vcovHC, type = "HC1")


#a. Yi = β0 + β1D1i + β2D2i + ui ;

bi_model_b1 <- lm(formula = our_DataSalary ~ our_DataExperience +
our_Datafact_edu +

                    (our_DataExperience * our_Datafact_edu), data = salary)


summary(bi_model_b1)



#_____________Завдання 2: Взаємодія між неперервною та бінарною
змінною_____________________


#Xi ! = досвід роботи особи i, неперервна змінна

##D1i - 1, якщо особа має вищу освіту, 0 якщо не має


# Yi = β0 + β1Xi + β2Di + ui


bi_model_a2 <- lm(our_DataSalary ~ log(our_DataExperience_num)  + our_DataEdu
+ (log(our_DataExperience_num) * log(our_Datafact_edu)), data = salary)


summary(bi_model_a2)


# Yi = β0 + β1Xi + β2Di + β3 (Xi × Di ) + ui

bi_model_b2 <- lm( our_DataSalary ~ our_DataExperience + our_Datafact_edu +

                    (our_DataExperience * our_Datafact_edu))


summary(bi_model_b2)


#c. Yi = β0 + β1Xi + β2 (Xi × Di ) + ui
```

```
#Є два варіанти взаємодії:
#Yi = β0 + β1Xi
#Yi = β0 + (β1 +β2 )Xi

bi_model_c <- lm(our_DataSalary ~ our_DataExperience +  (our_DataExperience *
our_Datafact_edu))
summary(bi_model_c)
# друк підсумків коефіцієнтів
coeftest(bi_model_c, vcov. = vcovHC, type = "HC1")

#d. Yi = β0 + β1 ln(Xi ) + β2 (ln(Xi ) × Di ) + ui

bi_model_d <- lm(our_DataSalary ~  log(our_DataExperience) +
log(our_DataExperience) * our_Datafact_edu)
summary(bi_model_d)
# друк підсумків коефіцієнтів
coeftest(bi_model_d, vcov. = vcovHC, type = "HC1")

#e. ln(Yi ) = β0 + β1 ln(Xi ) + β2 (ln(Xi ) × Di ) + ui

bi_model_e <- lm(log(our_DataSalary) ~  log(our_DataExperience) +
log(our_DataExperience) * our_Datafact_edu)
summary(bi_model_e)
# друк підсумків коефіцієнтів
coeftest(bi_model_e, vcov. = vcovHC, type = "HC1")

#_____________Повний статистичний аналіз найкращої моделі_____________

bi_model_coef <- lm(our_DataSalary ~ our_DataExperience + our_Datafact_edu +
(our_DataExperience * our_Datafact_edu), data = salary)
summary(bi_model_coef)
# друк підсумків коефіцієнтів
coeftest(bi_model_coef, vcov. = vcovHC, type = "HC1")
```

```
#довірчі інтервали для коефіцієнтів моделі;
confint(bi_model_coef)



#_____________Побудова розсіювань_____________


plot(our_DataExperience, our_DataSalary,
     pch = 20,
     col = 3,
     main = "")


#plot for model c.
plot(our_DataExperience, our_DataSalary,
     pch = 20,
     col = 1 + our_Datafact_edu,
     main = "")


#plot for model d.
plot(our_DataExperience, log(our_DataSalary),
     pch = 20,
     col = 1 + our_Datafact_edu,
     main = "")


#plot for model e.
plot(log(our_DataExperience), log(our_DataSalary),
     pch = 20,
     col = 1 + our_Datafact_edu,
     main = "")
#____________Передбачення даних_____________________
bi_model_prediction <- lm(our_DataSalary ~ our_DataExperience +
our_Datafact_edu + (our_DataExperience * our_Datafact_edu))
predict(bi_model_prediction, newdata = data.frame("our_DataExperience" = 0,
"our_Datafact_edu" = 0))
#t-статистика всіх коефіцієнтів для рівня значущості 5%;
coeftest(bi_model_prediction, vcov = vcovHC, type = "HC1")
```

```
confint(bi_model_prediction)
```

```
#__________________Завдання 3: Взаємодія між двома неперервними
змінними_____________
```

#$\boldsymbol{Y}$ — приріст заробітків і двома неперервними регресорами

#$\boldsymbol{X}$ **1** — роки стажу роботи

#$\boldsymbol{X}$ **2** — роки навчання.

#a. $Yi = \beta 0 + \beta 1 X 1 i + \beta 2 X 2 i + ui$

```
bi_model_a <- lm(our_DataSalary ~  our_DataExperience + our_DataEdu)
summary(bi_model_a)
# друк підсумків коефіцієнтів
coeftest(bi_model_a, vcov. = vcovHC, type = "HC1")
```

#b. $Yi = \beta 0 + \beta 1 X 1 i + \beta 2 X 2 i + \beta 3 \ (X 1 i \times X 2 i \ ) + ui$

```
bi_model_a <- lm(our_DataSalary ~  our_DataExperience + our_DataEdu +
(our_DataExperience * our_DataEdu))
summary(bi_model_a)
# друк підсумків коефіцієнтів
coeftest(bi_model_a, vcov. = vcovHC, type = "HC1")
```

#d. $Yi = \beta 0 + \beta 1 \ \ln(X 1 i \ ) + \beta 2 X 2 i + \beta 3 \ (\ln(X 1 i \ ) \times \ln(X 2 i \ )) + ui$

```
bi_model_danya <- lm(our_DataSalary ~ log(our_DataExperience) + our_DataEdu +
(log(our_DataExperience) * log(our_DataEdu)))
summary(bi_model_danya)
# друк підсумків коефіцієнтів
coeftest(bi_model_danya, vcov. = vcovHC, type = "HC1")
```

```
str(our_DataExperience)
```

```
str(our_DataEdu)

#_____________Повний статистичний аналіз найкращої моделі_____________

bi_model_coef <- lm(log( Salary) ~ Dnipro + Sex + Education+ log(Experience)
+   EXP_SQR +  Experience* Education + English + ( Sex* Experience*
Education), data=our_Data )


summary(bi_model_coef)

# друк підсумків коефіцієнтів

#довірчі інтервали для коефіцієнтів моделі;

confint(bi_model_coef)

#import даних

Training_set <- read_excel("Training set.xlsx")

View(Training_set)

#prediction using training set

help("predict")

prediction <- predict(bi_model_coef, newdata = Training_set)

df <- dplyr::tibble(exp(prediction))

View(df)
```