МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет
імені Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

ДОПУЩЕНО ДО ЗАХИСТУ
Завідувач кафедри інтелектуальних
інформаційних систем, д-р техн. наук, проф.
_____Ю. П. Кондратенко
«____»_____2024 р.

КВАЛІФІКАЦІЙНА РОБОТА БАКАЛАВРА

ІНФОРМАЦІЙНА СИСТЕМА АНАЛІЗУ І ПРОГНОЗУВАННЯ МЕДИЧНИХ ВИТРАТ

Спеціальність 122 «Комп'ютерні науки»

122 – КРБ – 402.22010226

Виконав студент 4-го курсу, групи 402
_____ Д. Д. Усов
«17» червня 2024 р.

Керівник: д-р. техн. наук, доцент
_____ І. О. Калініна
«17» червня 2024 р.

Миколаїв – 2024

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Чорноморський національний університет ім. Петра Могили
Факультет комп'ютерних наук
Кафедра інтелектуальних інформаційних систем

Рівень вищої освіти **бакалавр**
Спеціальність **122 «Комп'ютерні науки»**
(шифр і назва)
Галузь знань **12 «Інформаційні технології»**
(шифр і назва)

ЗАТВЕРДЖУЮ

Завідувач кафедри інтелектуальних інформаційних систем, д-р техн. наук, проф.
_____ Ю. П. Кондратенко
«____» _____2024 р.

З А В Д А Н Н Я
на виконання кваліфікаційної роботи

Видано студенту групи 402 факультету комп'ютерних наук Усову Данилу Дмитровичу.

1. Тема кваліфікаційної роботи «Інформаційна система аналізу і прогнозування медичних витрат».

Керівник роботи Калініна Ірина Олександрівна, кандидат техн. наук, доцент.

Затв. наказом Ректора ЧНУ ім. Петра Могили від «28» грудня 2023 р. № 271

2. Строк представлення кваліфікаційної роботи студентом «17» червня 2024 р.

3. Вхідні (початкові) дані до роботи: набір даних, що містить медичні витрати для пацієнтів.

Очікуваний результат: інформаційна система аналізу і прогнозування медичних витрат.

4. Перелік питань, що підлягають розробці (зміст пояснювальної записки):

- аналіз актуальності задачі про розробку інформаційної системи для страхових компаній;
- розгляд засобів програмної реалізації;

– програмна реалізація інформаційної системи аналізу і прогнозування медичних витрат;

– аналіз отриманих результатів роботи.

5. Перелік графічного матеріалу: 24 ілюстрації, 5 таблиць, презентація.

6. Завдання до спеціальної частини: «Визначення типу та параметрів штучної системи вентиляції для робочого приміщення»

7. Консультанти розділів роботи

Розділ	Прізвище, ініціали та посада консультанта	Підпис
Спеціальна частина з охорони праці	Алексєєва А.О., доцент кафедри екології	

Керівник роботи __ д-р техн. наук, доцент Калініна І. О._____
(наук. ступінь, вчене звання, прізвище та ініціали)

_____
(підпис)

Завдання прийнято до виконання ___Усов Д. Д._____
(прізвище та ініціали)

_____
(підпис)

Дата видачі завдання «_14_» _____січня____ 2024 р.

КАЛЕНДАРНИЙ ПЛАН
виконання кваліфікаційної роботи

Тема: Інформаційна система аналізу і прогнозування медичних витрат_____

№	Найменування роботи	Початок	Закінчення	Примітки
1	Подання заяви на затвердження теми та керівників КРБ	10.11.2023	15.11.2023	Виконано
2	Отримання завдання на виконання КРБ	10.01.2024	15.01.2024	Виконано
3	Складання календарного плану роботи на весь період виконання КРБ	16.01.2024	30.01.2024	Виконано
4	Отримання завдання на переддипломну практику	15.04.2024	29.04.2024	Виконано
5	Проходження переддипломної практики, збір та аналіз матеріалів до КРБ	29.04.2024	11.05.2024	Виконано
6	Розробка звіту з переддипломної практики	12.05.2024	15.05.2024	Виконано
7	Виконання КРБ аналіз сучасного стану задачі прогнозування медичних витрат, розробка ПЗ	13.05.2024	22.06.2024	Виконано
8	Перший попередній захист КРБ на засіданні комісії кафедри	27.05.2024	27.05.2024	Виконано
9	Доробка та остаточне оформлення КРБ	28.05.2024	09.06.2024	Виконано
10	Другий попередній захист КРБ на засіданні комісії кафедри	10.06.2024	10.06.2024	Виконано
11	Подання КРБ рецензенту	13.06.2024	13.06.2024	Виконано
11	Подання КРБ, її електронної копії та інших документів (відгуку, рецензії) до захисту	17.06.2024	21.06.2024	
12	Захист БКР перед екзаменаційною комісією (ЕК)	24.06.2024	28.06.2024	

Розробив студент __Усов Д. Д._____ _____
(прізвище, ім'я, по батькові студента) (*підпис*)

Керівник роботи __д-р техн. наук, проф. Калініна І. О. _____
(*посада, прізвище, ім'я, по батькові)* *(підпис)*

«_29_» _____01_____ 2024 р.

АНОТАЦІЯ

кваліфікаційної роботи студента групи 402 ЧНУ ім. Петра Могили
Усова Данила Дмитровича
Тема: «Інформаційна система аналізу і прогнозування медичних витрат»

Актуальність полягає у потребі адаптації страхових систем до нових умов. Зростання кількості постраждалих, збільшення медичних витрат та необхідність швидкого реагування на змінювані обставини потребують впровадження ефективних інформаційних систем. Такі системи повинні забезпечувати аналіз великих обсягів даних та подальше прогнозування для підтримки прийняття важливих рішень.

Об'єкт роботи – процес створення інформаційної системи, яка дозволить аналізувати та прогнозувати медичні витрати.

Предмет роботи – прогностичні регресійні моделі для прогнозування медичних витрат.

Метою кваліфікаційної роботи є створення інформаційної системи, яка дозволить аналізувати та прогнозувати медичні витрати.

Пояснювальна записка складається зі вступу, трьох розділів, висновків та додатку. У першому розділі розглядається застосування інформаційних систем аналізу і прогнозування у різних сферах, включаючи страхування та медичні витрати. У другому розділі досліджено методи створення моделей на основі лінійної регресії та їх порівняння з метою визначення найбільш точної. У третьому розділі виконано та описано прогнозування на основі кращих моделі, створеною у попередньому розділі.

В результаті розроблено інформаційну систему аналізу та прогнозування медичних витрат на мові R з використанням регресії.

Кваліфікаційна робота містить 59 сторінок, 24 рисунки, 5 таблиць, 25 використаних джерел та 1 додаток.

Ключові слова: інформаційна система, аналіз, прогнозування, страхування, медичні витрати.

ABSTRACT

qualification work of a student of group 402 of ChNU named after Petro Mohyla Black Sea National University

Usov Danylo

Topic: "Information system of analysis and forecasting of medical expenses"

The relevance lies in the need to adapt insurance systems to new conditions. The increase in the number of victims, the increase in medical costs, and the need to quickly respond to changing circumstances require the implementation of effective information systems. Such systems must provide analysis of large volumes of data and subsequent forecasting to support critical decision making.

The object of the study is the process of creating an information system that will allow analyzing and forecasting medical expenses.

The subject of the research is predictive regression models for forecasting medical expenses.

The purpose of the qualification work is to create an information system that will allow analyzing and forecasting medical expenses.

The explanatory note consists of an introduction, three sections, conclusions and an appendix. The first chapter examines the application of information systems of analysis and forecasting in various areas, including insurance and medical expenses. The second chapter examines the methods of creating models based on linear regression and their comparison in order to determine the most accurate one. In the third section, forecasting based on the best model created in the previous section is performed and described.

As a result, an Information system of analysis and forecasting of medical expenses was developed in the R language using regression.

The qualification work contains 59 pages, 24 figures, 5 tables, 25 used sources and 1 appendix.

Key words: information system, analysis, forecasting, insurance, medical expenses.

ЗМІСТ

ВСТУП

У період воєнних дій страхування набуває особливої важливості через підвищені ризики та невизначеність. Війна приносить значні збитки, включаючи руйнування майна, втрату життя та здоров'я, що робить страхування необхідним інструментом для захисту громадян та бізнесів.

Створення інформаційної системи аналізу і прогнозування медичних витрат є актуальним завданням в сучасних умовах. Необхідність точного прогнозування медичних витрат стає важливим фактором для забезпечення якісного медичного обслуговування і фінансової стабільності.

Актуальність полягає у потребі адаптації страхових систем до нових умов. Зростання кількості постраждалих, збільшення медичних витрат та необхідність швидкого реагування на змінювані обставини потребують впровадження ефективних інформаційних систем. Такі системи повинні забезпечувати аналіз великих обсягів даних та подальше прогнозування для підтримки прийняття важливих рішень.

Об’єкт дослідження – процес створення інформаційної системи, яка дозволить аналізувати та прогнозувати медичні витрати.

Предмет дослідження – прогностичні регресійні моделі для прогнозування медичних витрат.

Метою кваліфікаційної роботи є створення інформаційної системи, яка дозволить аналізувати та прогнозувати медичні витрати.

Завдання, що потрібно виконати для створення такої інформаційної системи:

– проаналізувати та обробити вхідні дані;
– створити моделі на основі вхідних даних;
– визначити та порівняти ефективність моделей;
– виконати прогнозування використовуючи кращу модель;
– проаналізувати отримані результати.

Подальша робота з іншими дослідженнями у сфері страхування, управління медичними витратами та впровадження сучасних інформаційних технологій буде мати позитивний вплив на якість надання страхових послуг. Тобто, звичайні люди зможуть отримувати більш підходящі та вигідніші для себе варіанти.

Таким чином, розробка інформаційної системи аналізу і прогнозування медичних витрат є актуальною і важливою задачею. Це сприятиме не лише покращенню якості страхових і медичних послуг, але й забезпеченню фінансової стабільності медичних закладів, страхових компаній та простих людей.

1 АНАЛІЗ ІНФОРМАЦІЙНИХ СИСТЕМ СТРАХОВИХ КОМПАНІЙ ТА ПОСТАНОВКА ЗАДАЧІ

1.1 Актуальність та огляд предметної області

1.1.1 Інформаційні системи

Інформаційні системи відіграють ключову роль у сучасному суспільстві, забезпечуючи зберігання, обробку та передачу інформації. Їх використання охоплює різні сфери життя, включаючи бізнес, науку, медицину, освіту та багато інших.

Інформаційна система – це сукупність взаємопов'язаних компонентів, які збирають, обробляють, зберігають і розповсюджують інформацію для підтримки прийняття рішень, координації та контролю в організації. Також, інформаційна системи складається з певних компонентів [1,2].

1. Апаратне забезпечення: фізичні пристрої, такі як комп'ютери, сервери, мережеве обладнання.
2. Програмне забезпечення: системне та прикладне програмне забезпечення, яке виконує конкретні функції.
3. Бази даних: структуровані набори даних, які зберігаються та керуються системою управління базами даних.
4. Люди: користувачі, які взаємодіють з інформаційною системою.
5. Процедури: інструкції та правила, які регулюють використання інформаційної системи.
6. Мережі: канали комунікації, які дозволяють передавати дані між різними компонентами системи.

Процеси в інформаційних системах включають різні етапи обробки даних, такі як збирання, зберігання, обробка, аналіз і передача інформації. Збирання даних означає збір інформації з різних джерел, як внутрішніх, так і зовнішніх. Це може бути автоматичний збір через сенсори, ручний ввід через інтерфейси або отримання даних з інших систем. Зберігання даних забезпечує можливість

безпечного збереження інформації у базах даних для подальшого використання. Системи управління базами даних відіграють ключову роль на цьому етапі. Обробка даних включає трансформацію зібраних даних у корисну інформацію, що може бути сортування, фільтрація, обчислення або агрегування даних. Аналіз даних забезпечує можливість виявлення тенденцій, закономірностей та інших корисних даних через використання аналітичних інструментів. Передача даних передбачає розповсюдження інформації між користувачами або іншими системами через мережі, такі, як: Інтернет, локальні мережі або інші канали комунікації.

Існує декілька типів інформаційних систем, кожна з яких має свої особливості та призначення.

1. Операційні системи обробки транзакцій: системи, які автоматизують рутинні та щоденні операції, такі як обробка замовлень, платежів та інвентаризація.
2. Системи управління базами даних: програмні засоби, які забезпечують створення, зберігання, модифікацію та витяг даних з баз даних.
3. Інформаційно-аналітичні системи: системи, які забезпечують надання інформації для підтримки управлінських рішень, надаючи звіти та підсумкову інформацію.
4. Системи підтримки прийняття рішень: системи, які допомагають у прийнятті складних рішень, використовуючи моделі аналізу та симуляції.
5. Експертні системи: системи, які імітують процес прийняття рішень експерта в конкретній області, використовуючи базу знань та правила виведення.
6. Інформаційні системи управління ресурсами підприємства: комплексні системи, які інтегрують всі аспекти бізнесу, включаючи планування, виробництво, продажі, маркетинг та фінанси.
7. Інформаційні системи управління відносинами з клієнтами: системи, які допомагають керувати взаємодією з клієнтами, забезпечуючи інструменти для аналізу та управління інформацією про клієнтів.

Інформаційні системи є невід'ємною частиною сучасного суспільства, забезпечуючи управління інформацією в різних сферах діяльності. Ці системи не

лише полегшують повсякденні операції, також підтримують прийняття важливих рішень.

Інформаційні системи можна класифікувати за різними критеріями, що допомагає краще розуміти їх функціональність та сферу застосування. За рівнем управління існують наступні системи.

1. Операційні системи обробки транзакцій: ці системи обробляють великий обсяг транзакцій і виконують рутинні завдання. Вони забезпечують швидку та надійну обробку даних, необхідних для виконання щоденних операцій підприємства.

2. Тактичні інформаційні системи: забезпечують підтримку менеджерів середнього рівня у процесі планування та контролю. Ці системи обробляють дані, що отримуються від операційних систем, та надають інформацію, необхідну для прийняття тактичних рішень.

3. Стратегічні інформаційні системи: спрямовані на підтримку вищого керівництва в процесі прийняття стратегічних рішень. Вони надають аналітичну інформацію та прогнози, необхідні для визначення довгострокової стратегії розвитку організації.

Також їх можна класифікувати за сферою застосування.

1. Бізнес-інформаційні системи: включають системи, які використовуються в різних аспектах бізнесу.

2. Інформаційні системи для науки та освіти: ці системи призначені для підтримки досліджень, зберігання наукових даних та забезпечення навчального процесу.

3. Медичні інформаційні системи: забезпечують зберігання та обробку медичних даних, управління пацієнтами та підтримку медичних процесів.

4. Інформаційні системи для державного управління: системи, що підтримують роботу державних установ, включаючи системи управління документами, реєстрами та інші.

Також, доцільно навести фрагмент блок-схеми інформаційної системи аналізу і прогнозування (рис. 1.1).

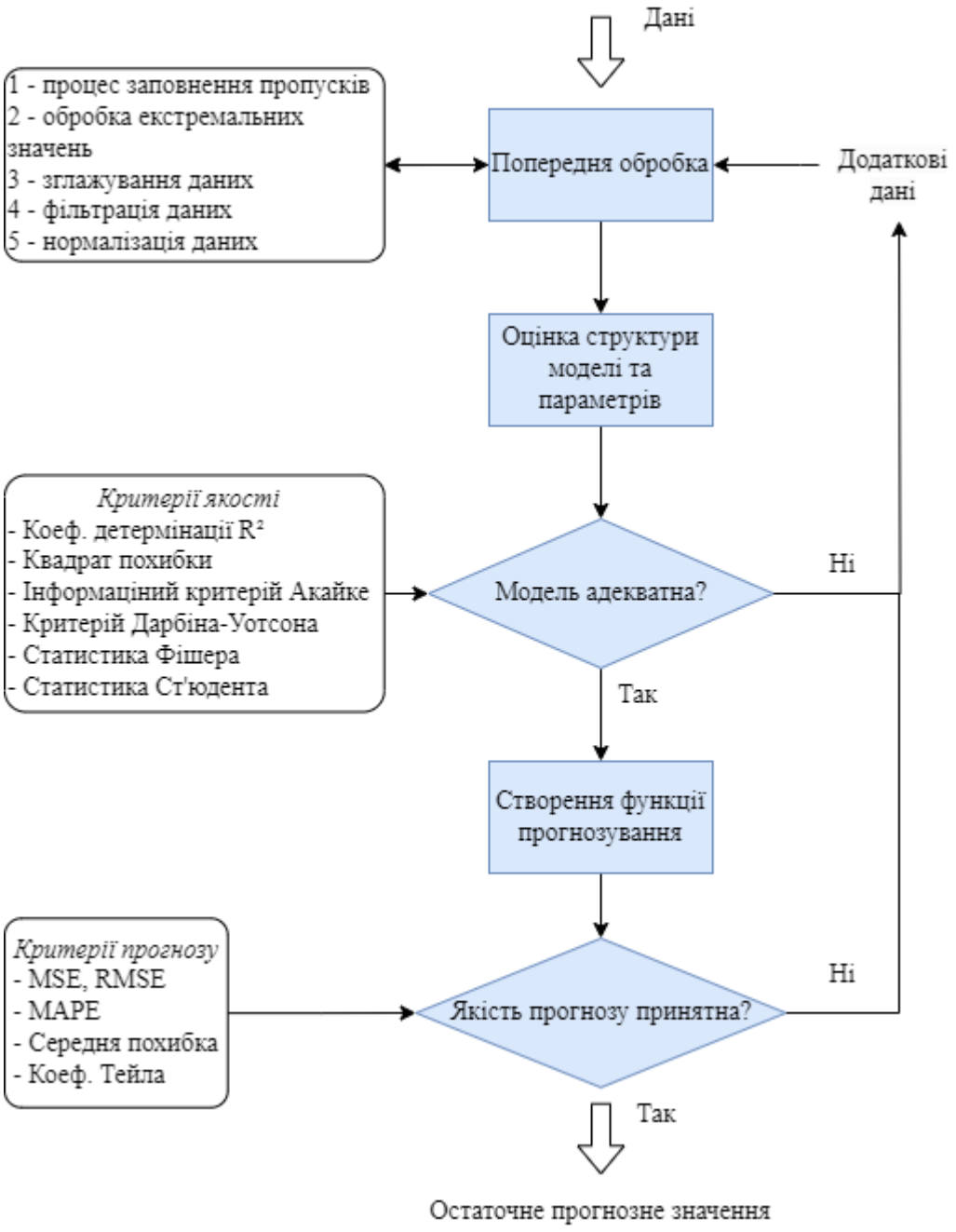

Рисунок 1.1 – Фрагмент блок-схеми інформаційної системи

1.1.2 Страхування

Страхування – це механізм захисту від фінансових втрат, який передбачає укладення договору між страхувальником (фізичною або юридичною особою) і страховиком (страховою компанією). У рамках цього договору страхувальник

сплачує страхові внески, а страховик зобов'язується компенсувати збитки у випадку настання певних ризиків.

Страхування є актуальним у наш час, за останні роки кількість нещасних випадків збільшилась, виплати страховиків зросли [3,4].

Таблиця 1.1 – Статистика нещасних випадків в Україні за останні роки

Рік	**Кількість нещасних випадків**	**Кількість постраждалих**	**Кількість випадків зі смертельним наслідком**
2020	8,000	8,500	500
2021	7,800	8,300	480
2022	7,600	8,100	460
2023	8,200	8,700	520

Таблиця 1.2 – Статистика страхових виплат в Україні за останні роки

Рік	**Страхові виплати, млрд грн**
2020	13,0
2021	13,5
2022	14,0
2023	16,2

У період воєнних дій страхування набуває особливої важливості через підвищені ризики та невизначеність. Війна приносить значні збитки, включаючи руйнування майна, втрату життя та здоров'я, що робить страхування необхідним інструментом для захисту громадян та бізнесів.

1. Захист майна: воєнні дії призводять до пошкодження або знищення житлових будинків, комерційних об'єктів та іншого майна. Майнове страхування допомагає відновити втрати та забезпечити фінансову стабільність.

2. Медичне страхування: в умовах підвищеного ризику травм та хвороб медичне страхування забезпечує доступ до необхідної медичної допомоги та покриває витрати на лікування.

3. Страхування життя: у воєнний час ризик втрати життя значно зростає. Страхування життя гарантує фінансову підтримку родини у разі загибелі страхувальника, забезпечуючи захист їхніх інтересів.

4. Бізнес-страхування: війна може призвести до значних збитків для бізнесу, включаючи втрату майна, зупинку виробництва та інші фінансові втрати. Страхування бізнесу допомагає зменшити ці ризики та швидше відновити діяльність після кризи.

Процес страхування включає кілька основних етапів. Оцінка ризиків є ключовим, під час якого страховик аналізує ризики, пов'язані з об'єктом страхування, щоб визначити розмір страхового внеску. Після оцінки ризиків сторони укладають договір, підписуючи страховий поліс, де зазначаються умови страхування, розмір внесків та обсяг покриття. Наступним етапом є сплата внесків, яку страхувальник здійснює регулярно відповідно до умов договору, забезпечуючи фінансову основу для виконання зобов'язань страховиком. У разі настання страхового випадку страховик оцінює збитки і виплачує страхове відшкодування згідно з умовами договору, що може включати грошову компенсацію або покриття витрат на відновлення майна чи здоров'я. Цей процес забезпечує фінансовий захист і стабільність для страхувальників, допомагаючи їм справлятися з непередбачуваними подіями.

З розвитком технологій інформаційні системи стали невід'ємною частиною страхового бізнесу. Вони забезпечують автоматизацію процесів, підвищення ефективності та покращення якості обслуговування клієнтів.

Інформаційні системи в страховій галузі забезпечують ефективне управління даними, автоматизацію бізнес-процесів та підвищення рівня обслуговування клієнтів. Основні види інформаційних систем, що використовуються у страхуванні наступні [5].

1. Системи управління відносинами з клієнтами: дозволяють зберігати та обробляти інформацію про клієнтів, управління взаємовідносинами, аналізувати потреби клієнтів та забезпечувати персоналізований підхід до обслуговування.

2. Системи управління страховими полісами: забезпечують автоматизацію процесів оформлення, обліку та адміністрування страхових полісів, що дозволяє значно скоротити час на обробку документів та зменшити ймовірність помилок.

3. Системи управління ризиками: допомагають аналізувати та оцінювати ризики, пов'язані зі страхуванням, прогнозувати можливі збитки та розробляти стратегії їхнього мінімізації.

4. Системи андеррайтингу: забезпечують автоматизацію процесу оцінки ризиків і визначення умов страхування, що дозволяє швидше та точніше приймати рішення про укладання договорів страхування.

5. Системи управління виплатами: автоматизують процес обробки страхових випадків, включаючи реєстрацію, оцінку збитків та виплату відшкодувань, що забезпечує швидке та ефективне вирішення питань клієнтів.

6. Аналітичні системи: використовуються для аналізу великих обсягів даних, виявлення тенденцій та закономірностей, що дозволяє покращити управлінські рішення та оптимізувати бізнес-процеси.

Для того, щоб медична страхова компанія могла функціонувати прибутково, необхідно, щоб сума щорічних страхових внесків перевищувала витрати на медичне обслуговування бенефіціарів. Це завдання є складним і вимагає значних зусиль, оскільки витрати на медичне обслуговування можуть бути непередбачуваними. Тому страхові компанії вкладають багато часу та грошей у розробку моделей, які точно прогнозують медичні витрати застрахованого населення.

Медичні витрати важко оцінити через те, що найдорожчі випадки трапляються рідко і здаються випадковими. Проте, існують певні закономірності, які можна врахувати. Наприклад, рак легень частіше зустрічається у курців, ніж у некурців, а від хвороб серця частіше страждають люди з надмірною вагою. Такі ситуації дозволяють визначити певні групи ризику серед населення.

Метою цього аналізу є використання даних про пацієнтів для прогнозування середніх витрат на медичне обслуговування для подібних груп населення. Це

завдання є важливим, оскільки такі оцінки можуть бути використані для створення страхових таблиць, які визначають суму щорічних внесків залежно від очікуваних витрат на лікування. Наприклад, для груп з високим ризиком розвитку певних захворювань сума внесків може бути встановлена вище, а для груп з низьким ризиком – нижче.

Такий підхід дозволяє страховим компаніям більш точно планувати свої бюджети та забезпечувати фінансову стабільність. Крім того, це допомагає встановлювати більш справедливі внески для різних категорій застрахованих, що сприяє загальному покращенню системи медичного страхування. Точне прогнозування медичних витрат є ключовим елементом у забезпеченні ефективної роботи страхової компанії, це дозволяє їй надавати відповідні послуги своїм клієнтам.

1.2 Вибір технологій

Для створення інформаційної системи аналізу і прогнозування медичних витрат було обрано мову програмування R.

R має значні можливості для здійснення статистичних аналізів, включаючи лінійну і нелінійну регресію, класичні статистичні тести, аналіз часових рядів (серій), кластерний аналіз і багато іншого. R легко розбудовується завдяки використанню додаткових функцій і пакетів [6].

Мова R є потужним інструментом для статистичного аналізу, моделювання даних та візуалізації. Основні причини вибору R наступні:

– широкий спектр статистичних та аналітичних методів. R вже має вбудований набір інструментів для розробки інформаційних систем аналізу і прогнозування;

– велика кількість інструментів для візуалізації даних. R має бібліотеки для візуалізації даних, такі як ggplot2, які дозволяють створювати графіки і діаграми, для більш наглядної демонстрації даних;

– R може обробити великі обсяги даних. Це є критично важливим при роботі з медичними записами;

– R поширюється безкоштовно, що означає відсутність ліцензійних витрат;

– наявність скриптів. R дозволяє зберігати та багаторазово використовувати написані алгоритми.

Також було обрано RStudio.

RStudio – вільне та відкрите інтегроване середовище розробки для R [7].

Воно дозволяє полегшити взаємодію з мовою програмування R. Основні причини вибору наступні:

– зручне середовище розробки, що включає: редактор коду з підсвіткою синтаксису, консоль для виконання команд R, панелі для перегляду графіків, змінних і файлів. Все це дозволяє ефективно працювати з кодом, та пришвидшує розробку інформаційних систем;

– RStudio підтримує створення та управління проєктами, що спрощує організацію файлів і коду. Також має функцію збереження зміних у середовищі.

1.3 Постановка задачі

У сучасному світі, особливо у воєнний час, страхування набуває ще більшої актуальності. Підвищений рівень ризиків та непередбачуваних обставин змушує страхові компанії адаптуватися до нових умов. Швидкий розвиток інформаційних технологій відкриває нові можливості для аналізу та прогнозування. Це є важливим для точного визначення страхових внесків. Для забезпечення фінансової стабільності та надання якісних послуг своїм клієнтам, страхові компанії потребують інформаційні системи, які допоможуть їм приймати обґрунтовані рішення.

Основні задачі, які треба розв’язати для створення такої інформаційної системи:

– проаналізувати та обробити вхідні дані;

– створити моделі на основі вхідних даних;

– визначити та порівняти ефективність моделей;
– виконати прогнозування використовуючи кращу модель;
– проаналізувати отримані результати.

1.4 Висновки до розділу

У першому розділі було розглянуто основні аспекти інформаційних систем та їх значення у сучасному суспільстві. Описано складові інформаційних систем, такі як апаратне та програмне забезпечення, бази даних, користувачі, процедури та мережі. Наведено типи інформаційних систем і їх класифікацію за рівнем управління та сферою застосування.

Окрему увагу приділено інформаційним системам у страховій галузі, які є невід'ємною частиною сучасного страхового бізнесу. Описано роль інформаційних систем у підвищенні ефективності управління даними, автоматизації процесів та покращенні обслуговування клієнтів. Розглянуто види інформаційних систем, що використовуються у страхуванні, та їх основні функції.

З розвитком технологій та зростанням обсягів даних, страхові компанії потребують сучасних інформаційних систем для точного прогнозування медичних витрат та визначення страхових внесків. Для цього було обрано мову програмування R та інтегроване середовище розробки RStudio, які забезпечують зручні інструменти для статистичного аналізу, моделювання даних та візуалізації.

2 СТВОРЕННЯ ПРОГНОСТИЧНИХ МОДЕЛЕЙ

2.1 Аналіз і попередня підготовка даних

Для аналізу було використано набір даних, що містить медичні витрати для пацієнтів, які проживають у Сполучених Штатах Америки. Дані було створено з використанням демографічної статистики, наданої Бюро перепису населення США, і, таким чином, приблизно відповідають реальним умовам.

У вибірці представлено 1338 бенефіціарів, які зареєстровані в програмі страхування, з ознаками, що відповідають характеристикам пацієнта, а також загальні медичні витрати, що входять до програми страхування за календарний рік.

Вибірка має наступні ознаки пацієнтів.

Таблиця 2.1 – Опис вихідних даних

Назва	Опис	Тип
age	Позначає вік основного бенефіціара (не старіше за 64 роки)	Числовий
sex	Стать страхувальника, чоловіча або жіноча	Категоріальний
bmi	Індекс маси тіла (ІМТ, або BMI), який дозволяє визначити, чи має людина недостатню або надмірну вагу. Ідеальний ІМТ знаходиться в межах від 18,5 до 24,9	Числовий

Закінчення таблиці 2.1

children	Позначає кількість дітей або утриманців, на яких поширюється програма страхування	Числовий
smoker	Приймає значення «так» або «ні» і вказує на те, чи курить застрахована особа	Категоріальний
region	Місце проживання одержувача страхування в США. Поділено на чотири географічні регіони: північний схід, південний схід, південний захід та північний захід	Категоріальний

Важливо врахувати те, яким чином оплачувані медичні витрати можуть залежати від цих змінних. Наприклад, можна очікувати, що літні люди і курці більш схильні до ризику великих медичних витрат. На відміну від багатьох інших методів машинного навчання, при регресійному аналізі залежність між ознаками зазвичай визначається користувачем, а не розпізнається автоматично.

Щоб завантажити дані для аналізу, було використано функцію read.csv() [8]. Також потрібно використати параметр stringsAsFactors=TRUE, тому що три номінальні змінні доцільно перетворити у фактори:

```
insurance <- read.csv("insurance.csv", stringsAsFactors = TRUE)
```

Функція str() підтверджує, що дані відформатовані так, як очікувалося:

```
> str(insurance)
```

```
'data.frame': 1338 obs. of 7 variables:
 $ age      : int  19 18 28 33 32 31 46 37 37 60 ...
 $ sex      : Factor w/ 2 levels "female","male": 1 2 2 2 2 1 ...
 $ bmi      : num  27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 ...
 $ children : int  0 1 3 0 0 0 1 3 2 0 ...
 $ smoker   : Factor w/ 2 levels "no","yes": 2 1 1 1 1 1 1 1 ...
 $ region   : Factor w/ 4 levels "northeast","northwest",..: ...
 $ expenses : num  16885 1726 4449 21984 3867 ...
```

Залежна змінна в даній моделі – це expenses, витрати на медичне обслуговування, які покриваються медичною страховкою протягом року для кожної людини. Перед тим як будувати регресійну модель, було перевірено нормальність даних. Для лінійної регресії залежна змінна не обов'язково повинна мати нормальний розподіл, однак часто модель виходить краще, коли ця умова виконується. Нижче наведено зведену статистику:

```
> summary(insurance$expenses)
   Min.  1st Qu.  Median    Mean   3rd Qu.   Max.
  1122    4740    9382    13270   16640   63770
```

Як можна побачити, середнє значення більше медіани. Це означає, що розподіл витрат на страхування має зсув вправо. Також це видно на рис. 2.1.

```
> hist(insurance$expenses)
```

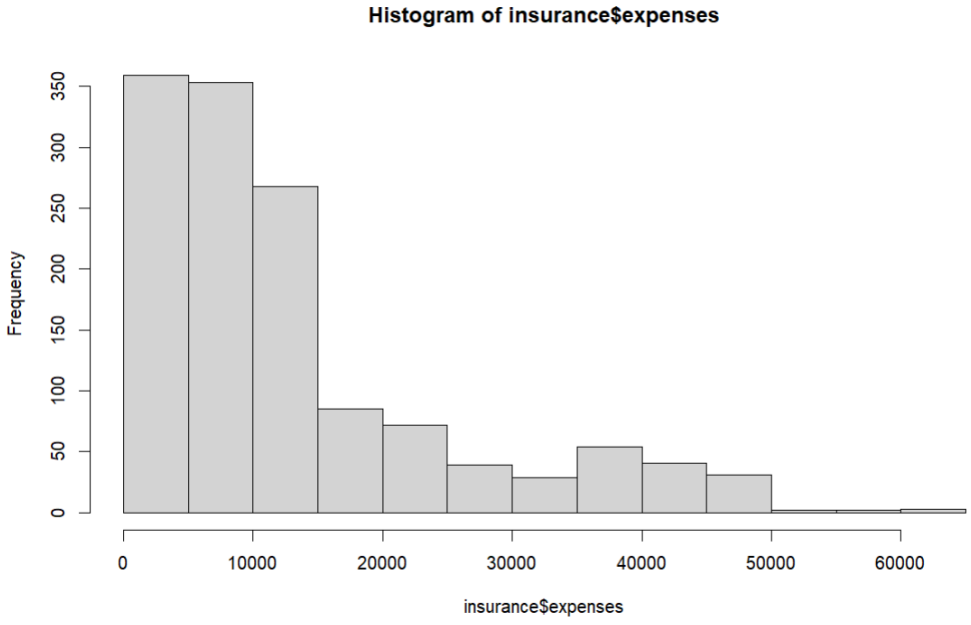

Рисунок 2.1 – Розподіл річних витрат на страхування

На рис. 2.1 видно, що розподіл зміщено вправо. Це говорить про те, що у більшості людей в наших даних щорічні медичні витрати становлять від 0 до 15 000 доларів, однак хвіст розподілу простягається далеко за межі цих пікових значень. Таке розподілення неідеально для лінійної регресії, однак знання про цей недолік, допоможе надалі розробити більш правильну модель.

Перш ніж вирішити цю проблему, потрібно звернути увагу на ще одну. Для регресійних моделей вимагається, щоб усі ознаки були числовими, однак у даному фреймі даних є три факторних елементи. Зокрема, змінна sex приймає значення male та female, а змінна smoker має категорії yes та no. З результатів summary() можна сказати, що ознака region має чотири рівні, тож було переглянуто, як вони розподіляються:

```
> table(insurance$region)
northeast northwest southeast southwest
      324       325       364       325
```

Як можна побачити, дані розподіляються майже порівну між чотирма географічними регіонами.

Перш ніж підлаштовувати регресійну модель під дані, було потрібно визначити, яким чином незалежні змінні пов'язані з залежною змінною і одна з одною. Швидку відповідь на це питання дає матриця кореляції. Вона показує кореляцію для кожної пари змінних з заданого набору [9].

Для того щоб створити матрицю кореляції для чотирьох числових змінних з фрейму даних про страхування, було використано команду cor():

```
> cor(insurance[c("age", "bmi", "children", "expenses")])
             age        bmi        children   expenses
age       1.0000000  0.1093410  0.0424690  0.2990082
bmi       0.1093410  1.0000000  0.0126447  0.1985763
children  0.0424690  0.0126447  1.0000000  0.0679982
expenses  0.2990082  0.1985763  0.0679982  1.0000000
```

На перетині кожного рядка і стовпця знаходиться кореляція для пари змінних, відповідних цьому рядку і стовпцю. Значення на діагоналі завжди дорівнює 1.0000000, оскільки кореляція змінної до самої себе завжди ідеальна.

Значення, розташовані симетрично відносно діагоналі, ідентичні, оскільки ці кореляції однакові. Іншими словами, cor(x, y) дорівнює cor(y, x).

Жодна з кореляцій у даній матриці не є дуже сильною, однак є кілька помітних залежностей. Зокрема, є слабка позитивна кореляція між age і bmi, що означає, що з віком маса тіла збільшується. Є також позитивна кореляція між ознаками age і expenses, bmi і expenses, а також children і expenses. Ці залежності означають, що з віком, збільшенням маси тіла і народженням дітей очікувана вартість страхування зростає.

Також було корисно візуалізувати відносини між числовими об'єктами за допомогою діаграм розсіювання. Можна було б побудувати діаграму розсіювання для кожної можливої пари змінних, однак такий підхід не є раціональним при великій кількості ознак.

Замість цього було побудовано матрицю розсіювання (рис. 2.2), яка представляє собою набір діаграм розсіювання, представлених у вигляді сітки. Матриця розсіювання застосовується для виявлення закономірностей серед трьох і більше змінних [10]. Вона не є справжньою багатовимірною візуалізацією, оскільки одночасно розглядаються лише дві ознаки. Однак дає загальне уявлення про взаємозв'язки між даними.

Для побудови матриці розсіювання для чотирьох числових ознак: age, bmi, children і expenses, було використано графічні можливості R. Функція pairs(), що входить до складу стандартного пакету R, надає базові функціональні можливості для побудови матриць розсіювання. Для того щоб викликати цю функцію, їй було надано фрейм даних для побудови діаграм. З фрейму даних про страхування було обрано тільки чотири числові змінні:

```
> pairs(insurance[c("age", "bmi", "children", "expenses")])
```

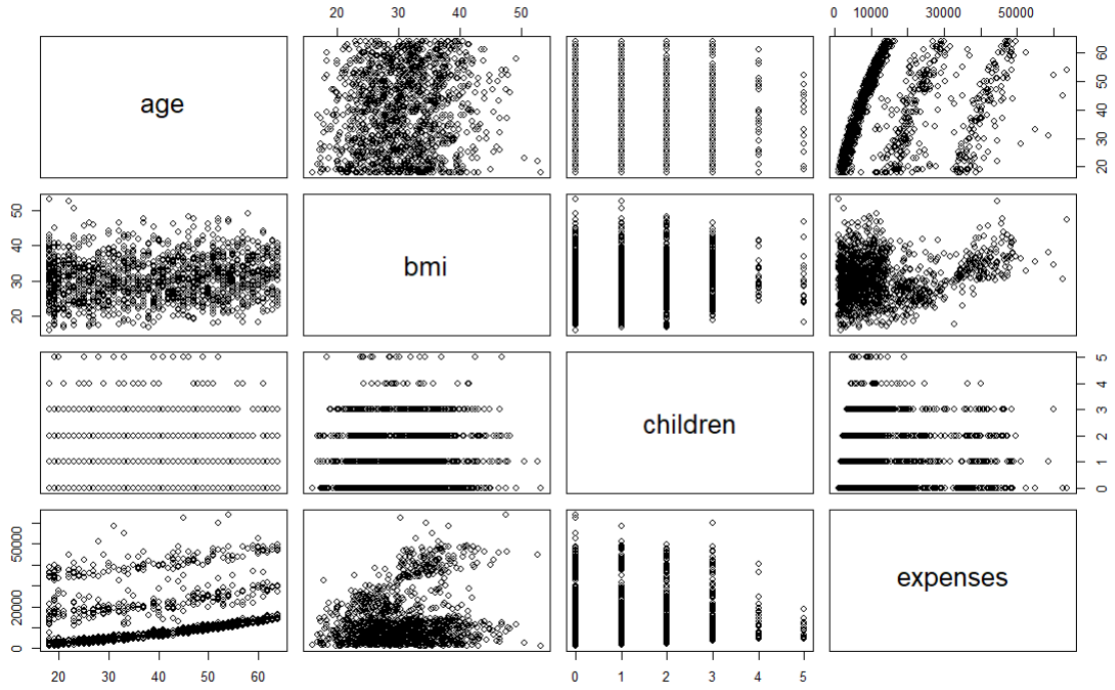

Рисунок 2.2 – Матриця розсіювання для числових ознак

Деякі закономірності виглядають як випадкові хмари точок, однак на інших, помітні певні тенденції. Взаємозв'язок між змінними age і expenses представлений кількома відносно прямими лініями, також у залежності expenses від bmi можна побачити дві групи точок. На інших діаграмах важко виявити які-небудь тенденції.

Потрібно було додати до діаграм більше інформації, щоб зробити їх більш корисними. За допомогою функції pairs.panels() з пакета psych було побудовано розширену матрицю розсіювання:

```
> pairs.panels(insurance[c("age", "bmi", "children", "expenses")])
```

У результатах діаграми розсіювання на рис. 2.3, представлені вище діагоналі, замінюються матрицями кореляції. Тепер на діагоналі містяться гістограми, що відображають розподіл значень для кожної ознаки. Також, діаграми розсіювання, розташовані нижче діагоналі, представлені з додатковою візуальною інформацією.

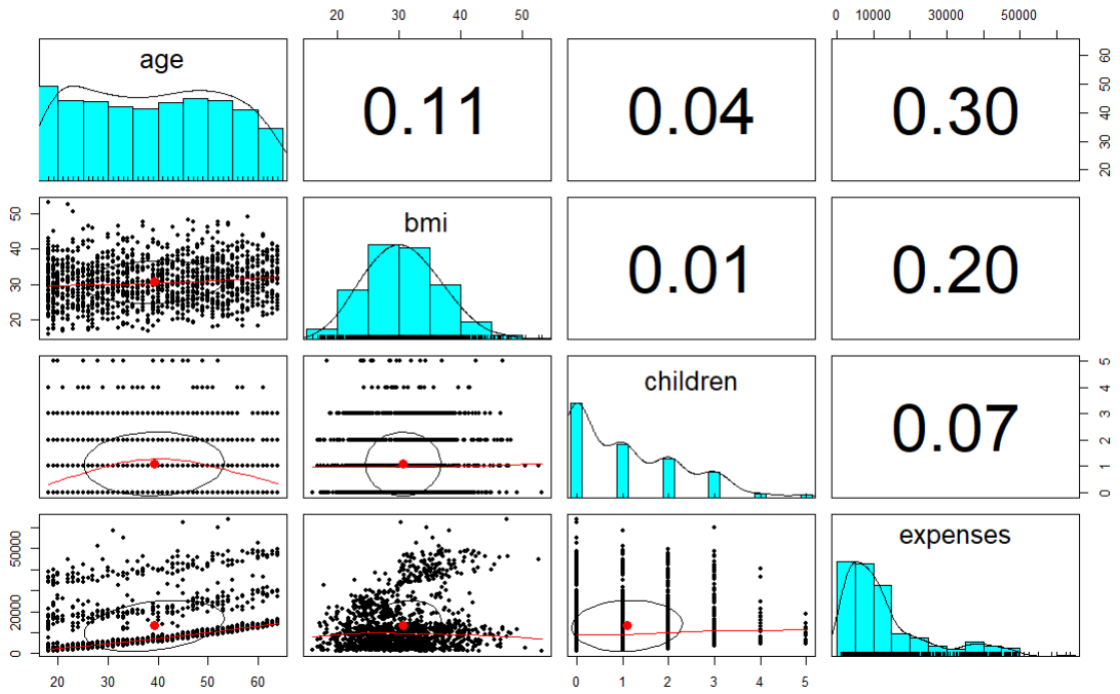

Рисунок 2.3 – Більш інформативна матриця розсіювання

Об'єкт овальної форми, присутній на кожній діаграмі розсіювання – це еліпс кореляції. Його призначення – візуалізація сили кореляції. Чим сильніше розтягнутий еліпс, тим сильніше кореляція. Майже ідеальне коло, як у залежності між bmi і children, вказує на дуже слабку кореляцію (0,01).

Еліпс для залежності між age і expenses витягнутий більше, що говорить про сильнішу кореляцію (0,30). Точка в центрі еліпса відображає середні значення змінних по осях X і Y.

Криві, зображені на діаграмах розсіювання – це LOESS-криві. Вони вказують на загальну залежність між змінними осей X і Y [11]. Крива для змінних age і children має сильно виражений пік відповідний середньому віку. Це означає, що найстаріші та наймолодші люди в даній вибірці мають менше дітей, які враховуються у програмі страхування, ніж люди середнього віку. Оскільки ця тенденція є нелінійною, то цей висновок не можна було зробити лише на підставі кореляцій. LOESS-крива для age і bmi, навпаки, являє собою пряму з незначним нахилом, що означає, що з віком маса тіла збільшується.

2.2 Створення моделей

Щоб підібрати модель лінійної регресії для даних за допомогою R, було використано функцію lm() [12]. Вона входить до складу пакета stats, який входить до стандартного пакета R і завантажується за замовчуванням.

За допомогою наступної команди було побудовано модель лінійної регресії, яка встановлює залежність сумарних медичних витрат від шести незалежних змінних. Вказувати величину зсуву для регресійної моделі немає необхідності, так як він враховується за замовчуванням:

```
model_1 <- lm(expenses ~ age + children + bmi + sex + smoker + region,
data = insurance)
```

Оскільки для вказівки всіх ознак (крім тих, що вже вказані у формулі) може використовуватися символ крапки, наступна команда еквівалентна попередній:

```
> model_1 <- lm(expenses ~ ., data = insurance)
```

Після того як модель побудована, було введено ім'я об'єкта моделі, щоб побачити отримані бета-коефіцієнти (рис. 2.4) [13].

```
> model_1

Call:
lm(formula = expenses ~ ., data = insurance)

Coefficients:
    (Intercept)              age          sexmale              bmi         children
smokeryes  regionnorthwest
      -11941.6            256.8           -131.4            339.3            475.7
23847.5           -352.8
regionsoutheast  regionsouthwest
        -1035.6           -959.3
```

Рисунок 2.4 – Отримані бета-коефіцієнти

Зсув (intercept) – це прогнозоване значення витрат, коли незалежні змінні дорівнюють нулю. Однак у багатьох випадках зсув сам по собі має мало сенсу, оскільки зазвичай не буває ситуації, при якій всі ознаки дорівнюють нулю. У даному випадку не існує людей з нульовим віком і нульовим коефіцієнтом ІМТ. Отже, у зсуву немає інтерпретації для реального світу.

Бета-коефіцієнти вказують на передбачуване збільшення витрат при збільшенні кожної ознаки на одиницю за умови, що всі інші значення залишаються постійними. Наприклад, можна очікувати, що з кожним роком медичні витрати в середньому будуть збільшуватися на 256,8 долара за умови, що всі інші змінні не змінюються. Аналогічно, поява кожної дитини призводить до додаткових медичних витрат, які становлять в середньому 475,7 долара на рік, а збільшення ІМТ на одиницю викликає збільшення щорічних витрат на медичне обслуговування в середньому на 339,3 долара за інших рівних умов.

Варто зауважити, що, хоча у формулі моделі вказано лише шість ознак, у результатах, крім зсуву, зазначено ще вісім коефіцієнтів. Так виходить тому, що функція lm() автоматично застосовує фіктивне кодування до всіх включених у модель змінних, які мають тип фактора.

Фіктивне кодування дозволяє розглядати номінальну ознаку як числову, створюючи двійкову змінну для кожної категорії цієї ознаки [14]. Фіктивна змінна приймає значення 1, якщо спостереження потрапляє в зазначену категорію, і 0 – в іншому випадку. Наприклад, ознака sex має дві категорії: male і female. Її було розбито на дві двійкові змінні, яким R було присвоєно імена sexmale і sexfemale. Для спостережень, де sex=male, змінна sexmale=1, а sexfemale=0. І навпаки, якщо sex=female, то sexmale=0 і sexfemale=1. Таке ж саме кодування застосовується до змінних з трьома і більше категоріями. Таким чином, було розділено ознаку region з чотирма категоріями на чотири фіктивні змінні: regionnorthwest, regionsoutheast, regionsouthwest і regionnortheast.

При додаванні фіктивної змінної до регресійної моделі одна категорія завжди є еталонною. Решта оцінок інтерпретуються відносно еталонної. У даній моделі автоматично створені змінні sexfemale, smokerno і regionnortheast, таким чином, що некурящі жінки, які проживають у північно-східному регіоні, стають еталонною групою. Відповідно, у чоловіків щорічні медичні витрати на 131,4 долара менше, ніж у жінок, а курці витрачають на медицину в середньому на 23847,5 долара на рік більше, ніж некурящі. Для всіх інших трьох регіонів у цій моделі коефіцієнт є

від'ємним. Це означає, що еталонна група – північно-східний регіон – має в середньому найвищі витрати на медицину.

За замовчуванням у R в якості еталонного стає перший з рівнів факторної змінної.

Оцінки параметрів, які було отримано після виконання команди model_1, показують, як незалежні змінні пов'язані з залежною змінною, але не показують, наскільки добре модель відповідає даним. Щоб оцінити ефективність моделі, було використано команду summary() для збереженої моделі [15].

```
> summary(model_1)
```

На рис. 2.5 наявні три блоки, що відповідають трьом способам оцінити ефективність даної моделі, тобто її відповідність даним.

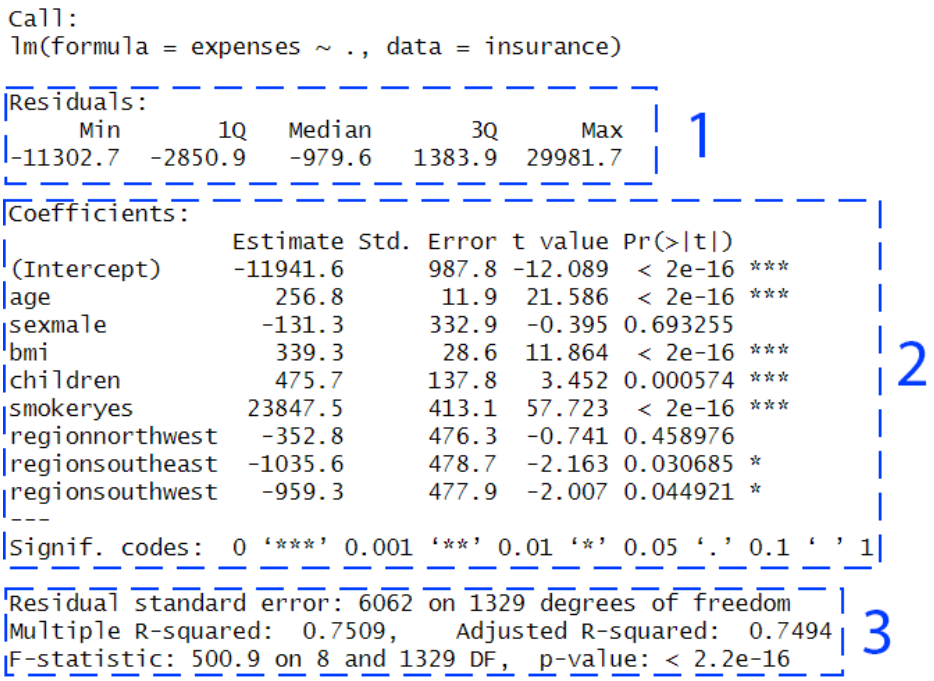

Рисунок 2.5 – Результат використання команди summary()

1. У розділі Residuals представлена зведена статистика помилок прогнозів, деякі з цих помилок досить значні. Оскільки залишкове значення дорівнює різниці між істинним і прогнозованим значеннями, максимальна помилка становить 29981,7 долари. Можна зробити припущення, що принаймні для одного спостереження моделлю було занижено прогнозовані витрати майже на 30000

доларів. З іншого боку, 50% помилок знаходяться в межах від 2828,7 долара більше істинного значення до 982,1 долара менше істинного значення.

2. Для кожного обчисленого коефіцієнта регресії визначено p-значення, позначене як Pr(>|t|), яке представляє собою оцінку ймовірності того, що для даної оцінки істинний коефіцієнт дорівнює нулю. Невеликі p-значення означають, що істинний коефіцієнт навряд чи дорівнює нулю, отже, малоймовірно, що дана ознака не пов'язана із залежною змінною. Деякі з p-значень позначені зірочками «***», тобто мають примітки. Ці примітки вказують на рівень значущості для даної оцінки. Рівень значущості є порогом, обраним до побудови моделі, який буде використовуватися для позначення «реальних», а не випадкових результатів. p-значення нижче рівня значущості вважаються статистично значущими. Якщо в моделі мало таких значень, це може бути приводом для занепокоєння, оскільки вказує на те, що використовувані ознаки не дуже добре прогнозують результат. У даній моделі є кілька досить значущих змінних, і вони логічно пов'язані з результатом.

3. Коефіцієнт детермінації дозволяє оцінити, наскільки добре модель в цілому пояснює значення залежної змінної. Ця величина схожа на коефіцієнт кореляції так, що чим ближче її значення до 1,0, тим краще модель пояснює дані. Оскільки в моделі коефіцієнт детермінації дорівнює 0,7494, це означає, що модель пояснює майже 75 % змін залежної змінної. Чим більше в моделі ознак, тим краще вони пояснюють зміни залежної змінної, коефіцієнт детермінації представляє собою значення, скориговане за рахунок накладання штрафів на моделі з великою кількістю незалежних змінних. Це корисно для порівняння ефективності моделей з різною кількістю пояснюючих змінних.

На підставі описаних трьох показників ефективності можна сказати, що отримана модель працює досить добре. Нерідко регресійні моделі реальних даних мають досить низький коефіцієнт детермінації: значення 0,75 – досить хороший результат. Величина деяких помилок може трохи турбувати, але це нормально, враховуючи характер даних про медичні витрати.

Головна відмінність регресійного моделювання від інших методів машинного навчання полягає в тому, що регресія зазвичай залишає користувачу вибір ознак і специфікацію моделі [16]. Отже, якщо є знання про те, як ознака пов'язана з результатом, то можна використати цю інформацію у специфікації моделі і підвищити її ефективність.

У лінійній регресії припускається, що зв'язок між незалежною і залежною змінними є лінійним [17]. Однак це не завжди так, наприклад, вплив віку на медичні витрати не може бути постійним для всіх вікових значень. Для літніх груп населення лікування може стати занадто дорогим [18].

Типове рівняння регресії має наступний вигляд:

$$y = \alpha + \beta_1 x.$$

Щоб врахувати нелінійні залежності, у рівняння регресії можна додати член вищого порядку, представивши модель у вигляді полінома. В цілому, було змодельовано наступні залежності:

$$y = \alpha + \beta_1 x + \beta_z x^2.$$

Різниця між цими двома моделями полягає в тому, що вводиться додатковий бета-коефіцієнт, призначений для відображення впливу члена x^2. Це дозволяє представити вплив віку як функцію квадрата віку.

Для того щоб додати в модель нелінійну залежність від віку, просто було створено ще одну змінну:

```
> insurance$age2 <- insurance$age^2
```

Припускається, що вплив ознаки не є накопичувальним, скоріше за все, він починає позначатися тільки після досягнення певного порогового значення. Наприклад, ІМТ може ніяк не впливати на медичні витрати для людей з нормальною вагою, але сильно впливати на більш високі витрати для огрядних людей (тобто для ІМТ зі значеннями 30 і більше).

Таким чином було змодельовано цей вплив, створено бінарну змінну індикатора ожиріння, яка дорівнює 1, якщо ІМТ перевищує 30, і 0, якщо не перевищує. Тоді оціночний бета-коефіцієнт для цієї бінарної ознаки вказує на

середній чистий вплив на медичні витрати для тих осіб, чий ІМТ дорівнює або перевищує 30, відносно тих, у кого ІМТ менше 30.

Щоб створити ознаку, було використано функцією ifelse(), яка перевіряє зазначену умову для кожного елемента вектора і повертає значення в залежності від того, є ця умова істинною або хибною. Для ІМТ, більше або рівного 30, повертається 1, в іншому випадку – 0:

```
> insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
```

Далі в покращену модель до змінної bmi було додано змінну bmi30.

До цього часу було розглянуто тільки індивідуальний внесок кожної ознаки в результат. Але може бути так, що окремі ознаки чинять спільний вплив на залежну змінну. Наприклад, куріння і ожиріння можуть чинити шкідливий вплив окремо, але розумно припустити, що їх сукупний ефект може виявитися гіршим, ніж від кожного з них окремо.

Якщо дві ознаки мають сукупний ефект – це є взаємодією. Якщо є підозра, що дві змінні взаємодіють, можна перевірити цю гіпотезу, додавши їх взаємодію в модель. Щоб задати взаємодію між індикатором ожиріння (bmi30) і індикатором куріння (smoker), було використано наступну формулу: expenses ~ bmi30 * smoker.

Оператор «*» – це скорочення, що вказує на те, що модель треба представити як expenses ~ bmi30 + smokeryes + bmi30:smokeryes. Оператор двокрапка «:» у розгорнутому вигляді вказує на те, що bmi30:smokeryes представляє собою взаємодію між двома змінними.

Після додавання змін у дані, які описано вище, було створено ще моделі. Наступним чином було створено модель 2.

```
> model_2 <- lm(expenses ~ age + age2 + children + bmi + sex + smoker
+ region, data = insurance)
```

Було додано нелінійну змінну для віку.

Як видно на рис. 2.7, значення r-squared дорівнює 0,7537 – це означає, що апроксимація моделі є хорошою [19]. Модель пояснює 75,37% мінливості змінної expenses. Також на рис. 2.6 наведено бета-коефіцієнти.

```
> model_2
```

```
Call:
lm(formula = expenses ~ age + age2 + children + bmi + sex + smoker +
    region, data = insurance)

Coefficients:
    (Intercept)             age            age2        children             bmi
      sexmale       smokeryes  regionnorthwest
      -6602.064         -54.423           3.925         642.121         335.291
      -138.451       23858.690        -367.632
regionsoutheast  regionsouthwest
      -1031.998         -956.787
```

Рисунок 2.6 – Отримані бета-коефіцієнти

```
> summary(model_2)

Call:
lm(formula = expenses ~ age + age2 + children + bmi + sex + smoker +
    region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11665.6  -2854.7   -942.7   1300.8  30814.6

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -6602.064   1689.528  -3.908 9.79e-05 ***
age               -54.423     80.989  -0.672 0.501716
age2                3.925      1.010   3.885 0.000107 ***
children          642.121    143.613   4.471 8.44e-06 ***
bmi               335.291     28.467  11.778  < 2e-16 ***
sexmale          -138.451    331.189  -0.418 0.675983
smokeryes       23858.690    410.976  58.054  < 2e-16 ***
regionnorthwest  -367.632    473.771  -0.776 0.437905
regionsoutheast -1031.998    476.164  -2.167 0.030388 *
regionsouthwest  -956.787    475.398  -2.013 0.044358 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6030 on 1328 degrees of freedom
Multiple R-squared:  0.7537,    Adjusted R-squared:  0.7521
F-statistic: 451.6 on 9 and 1328 DF,  p-value: < 2.2e-16
```

Рисунок 2.7 – Результат використання команди summary()

Наступним чином було створено модель 3.

```
> model_3 <- lm(expenses ~ age + children + bmi30*bmi + sex + smoker +
region, data = insurance)
```

Було введено залежність між змінними bmi та bmi30.

Як видно на рис. 2.9, значення r-squared дорівнює 0,756 – це означає, що апроксимація моделі є хорошою, а модель пояснює 75,6% мінливості змінної expenses. Також на рис. 2.8 наведено бета-коефіцієнти.

```
> model_3
```

```
Call:
lm(formula = expenses ~ age + children + bmi30 * bmi + sex +
    smoker + region, data = insurance)

Coefficients:
    (Intercept)              age         children            bmi30              bmi          sexmale
smokeryes  regionnorthwest
        -8714.4            257.0            477.4           4739.2            191.4           -162.7
23853.1          -409.5
regionsoutheast  regionsouthwest        bmi30:bmi
         -877.5           -966.0            -65.2
```

Рисунок 2.8 – Отримані бета-коефіцієнти

```
> summary(model_3)

Call:
lm(formula = expenses ~ age + children + bmi30 * bmi + sex +
    smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-11647.3  -3467.3   -220.4   1624.1  28356.2

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -8714.36    1999.36  -4.359 1.41e-05 ***
age               256.95      11.79  21.792  < 2e-16 ***
children          477.43     136.50   3.498 0.000485 ***
bmi30            4739.16    2796.50   1.695 0.090372 .
bmi               191.38      76.73   2.494 0.012746 *
sexmale          -162.69     329.85  -0.493 0.621931
smokeryes       23853.13     409.34  58.272  < 2e-16 ***
regionnorthwest  -409.54     472.82  -0.866 0.386553
regionsoutheast  -877.48     475.18  -1.847 0.065027 .
regionsouthwest  -965.99     474.03  -2.038 0.041767 *
bmi30:bmi         -65.20      94.77  -0.688 0.491620
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6005 on 1327 degrees of freedom
Multiple R-squared:  0.756,     Adjusted R-squared:  0.7541
F-statistic: 411.1 on 10 and 1327 DF,  p-value: < 2.2e-16
```

Рисунок 2.9 – Результат використання команди summary()

Наступним чином було створено модель 4.

```
> model_4 <- lm(expenses ~ age + age2 + children + bmi30*bmi + sex + smoker + region, data = insurance)
```

Було внесено такі зміни:

– додано нелінійну змінну для віку;

– введено залежність між змінними bmi та bmi30.

Як видно на рис. 2.11, значення r-squared дорівнює 0,7584 – це означає, що апроксимація моделі є хорошою, а модель пояснює 75,84% мінливості змінної expenses. Також на рис. 2.10 наведено бета-коефіцієнти.

```
> model_4

Call:
lm(formula = expenses ~ age + age2 + children + bmi30 * bmi + 
    sex + smoker + region, data = insurance)

Coefficients:
    (Intercept)              age             age2         children           bmi30              bmi  
sexmale        smokeryes  
      -4152.079          -30.882            3.629          631.144         4935.972          203.468  
-168.194        23864.628  
regionnorthwest  regionsoutheast  regionsouthwest        bmi30:bmi  
       -425.309         -879.512         -967.003          -76.425
```

Рисунок 2.10 – Отримані бета-коефіцієнти

```
> summary(model_4)

Call:
lm(formula = expenses ~ age + age2 + children + bmi30 * bmi + 
    sex + smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-12117.1  -3362.6     -6.2   1445.0  29171.0 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -4152.079   2356.944  -1.762 0.078361 .  
age               -30.882     80.508  -0.384 0.701347    
age2                3.629      1.004   3.614 0.000313 ***
children          631.144    142.386   4.433 1.01e-05 ***
bmi30            4935.972   2784.410   1.773 0.076505 .  
bmi               203.468     76.457   2.661 0.007880 ** 
sexmale          -168.194    328.364  -0.512 0.608583    
smokeryes       23864.628    407.509  58.562  < 2e-16 ***
regionnorthwest  -425.309    470.702  -0.904 0.366391    
regionsoutheast  -879.512    473.039  -1.859 0.063209 .  
regionsouthwest  -967.003    471.893  -2.049 0.040639 *  
bmi30:bmi         -76.425     94.394  -0.810 0.418294    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5978 on 1326 degrees of freedom
Multiple R-squared:  0.7584,    Adjusted R-squared:  0.7564 
F-statistic: 378.3 on 11 and 1326 DF,  p-value: < 2.2e-16
```

Рисунок 2.11 – Результат використання команди summary()

Наступним чином було створено модель 5.

```
> model_5 <- lm(expenses ~ age + bmi30*children + bmi + sex + smoker + region, data = insurance)
```

Було введено залежність між ожирінням і кількістю дітей.

Як видно на рис. 2.13, значення r-squared дорівнює 0,756 – це означає, що апроксимація моделі є хорошою, а модель пояснює 75,6% мінливості змінної expenses. Також на рис. 2.12 наведено бета-коефіцієнти.

```
> model_5
```

```
Call:
lm(formula = expenses ~ age + bmi30 * children + bmi + sex +
    smoker + region, data = insurance)

Coefficients:
    (Intercept)              age            bmi30         children              bmi          sexmale
smokeryes  regionnorthwest
        -7547.0            257.1           2644.1            380.2            149.3           -155.4
23848.7           -398.6
regionsoutheast  regionsouthwest   bmi30:children
         -881.9           -953.3            189.8
```

Рисунок 2.12 – Отримані бета-коефіцієнти

```
> summary(model_5)

Call:
lm(formula = expenses ~ age + bmi30 * children + bmi + sex +
    smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-12021.6  -3464.7   -116.8   1553.2  28578.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -7546.97    1289.55  -5.852 6.10e-09 ***
age               257.07      11.79  21.808  < 2e-16 ***
bmi30            2644.14     625.94   4.224 2.56e-05 ***
children          380.21     195.83   1.942  0.05241 .
bmi               149.26      46.27   3.226  0.00129 **
sexmale          -155.45     329.94  -0.471  0.63762
smokeryes       23848.71     409.25  58.275  < 2e-16 ***
regionnorthwest  -398.56     472.04  -0.844  0.39864
regionsoutheast  -881.92     475.07  -1.856  0.06362 .
regionsouthwest  -953.26     473.44  -2.013  0.04427 *
bmi30:children    189.75     273.00   0.695  0.48714
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6005 on 1327 degrees of freedom
Multiple R-squared:  0.756,     Adjusted R-squared:  0.7541
F-statistic: 411.1 on 10 and 1327 DF,  p-value: < 2.2e-16
```

Рисунок 2.13 – Результат використання команди summary()

Наступним чином було створено модель 6.

```
> model_6 <- lm(expenses ~ age + age2 + bmi30*children + bmi + sex + smoker + region, data = insurance)
```

Було внесено такі зміни:

– додано нелінійну змінну для віку;

– введено залежність між ожирінням і кількістю дітей.

Як видно на рис. 2.15, значення r-squared дорівнює 0,7583 – це означає, що апроксимація моделі є хорошою, а модель пояснює 75,83% мінливості змінної expenses. Також на рис. 2.14 наведено бета-коефіцієнти.

```
> model_6

Call:
lm(formula = expenses ~ age + age2 + bmi30 * children + bmi +
    sex + smoker + region, data = insurance)

Coefficients:
    (Intercept)              age             age2            bmi30         children              bmi
sexmale        smokeryes
      -2862.504          -27.697            3.591         2536.372          541.257          154.046
-161.135        23858.818
regionnorthwest  regionsoutheast  regionsouthwest   bmi30:children
       -409.626         -885.622         -951.035          172.485
```

Рисунок 2.14 – Отримані бета-коефіцієнти

```
> summary(model_6)

Call:
lm(formula = expenses ~ age + age2 + bmi30 * children + bmi +
    sex + smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-12418.1  -3366.1    135.9   1310.3  29398.7

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -2862.504   1833.953  -1.561 0.118800
age               -27.697     80.471  -0.344 0.730761
age2                3.591      1.004   3.577 0.000360 ***
bmi30            2536.372    623.901   4.065 5.08e-05 ***
children          541.257    200.098   2.705 0.006919 **
bmi               154.046     46.083   3.343 0.000852 ***
sexmale          -161.135    328.488  -0.491 0.623836
smokeryes       23858.818    407.450  58.556  < 2e-16 ***
regionnorthwest  -409.626    469.966  -0.872 0.383580
regionsoutheast  -885.622    472.971  -1.872 0.061362 .
regionsouthwest  -951.035    471.348  -2.018 0.043825 *
bmi30:children    172.485    271.840   0.635 0.525858
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5978 on 1326 degrees of freedom
Multiple R-squared:  0.7583,    Adjusted R-squared:  0.7563
F-statistic: 378.2 on 11 and 1326 DF,  p-value: < 2.2e-16
```

Рисунок 2.15 – Результат використання команди summary()

Наступним чином було створено модель 7.

```
> model_7 <- lm(expenses ~ age + bmi30*children + bmi + sex +
bmi30*smoker + region, data = insurance)
```

Було внесено такі зміни:

– введено залежність між ожирінням і кількістю дітей;

– введено залежність між ожирінням і курінням.

Як видно на рис. 2.17, значення r-squared дорівнює 0,8639 – це означає, що апроксимація моделі є хорошою, а модель пояснює 86,39% мінливості змінної expenses. Також на рис. 2.16 наведено бета-коефіцієнти.

```
> model_7

Call:
lm(formula = expenses ~ age + bmi30 * children + bmi + sex +
    bmi30 * smoker + region, data = insurance)

Coefficients:
    (Intercept)              age            bmi30         children              bmi          sexmale
smokeryes  regionnorthwest
        -4670.9            263.2           -995.3            458.4            114.8           -487.5
13405.5          -273.3
regionsoutheast  regionsouthwest   bmi30:children  bmi30:smokeryes
         -822.3          -1226.4            120.7          19790.5
```

Рисунок 2.16 – Отримані бета-коефіцієнти

```
> summary(model_7)

Call:
lm(formula = expenses ~ age + bmi30 * children + bmi + sex +
    bmi30 * smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-18169.7  -1848.9  -1252.0   -447.7  24804.5

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -4670.891    967.473  -4.828 1.54e-06 ***
age               263.163      8.808  29.876  < 2e-16 ***
bmi30            -995.265    480.906  -2.070 0.038687 *
children          458.399    146.322   3.133 0.001769 **
bmi               114.829     34.582   3.320 0.000923 ***
sexmale          -487.473    246.706  -1.976 0.048370 *
smokeryes       13405.529    444.060  30.189  < 2e-16 ***
regionnorthwest  -273.294    352.674  -0.775 0.438525
regionsoutheast  -822.317    354.919  -2.317 0.020660 *
regionsouthwest -1226.396    353.798  -3.466 0.000544 ***
bmi30:children    120.668    203.967   0.592 0.554214
bmi30:smokeryes 19790.493    610.292  32.428  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4486 on 1326 degrees of freedom
Multiple R-squared:  0.8639,    Adjusted R-squared:  0.8628
F-statistic: 765.2 on 11 and 1326 DF,  p-value: < 2.2e-16
```

Рисунок 2.17 – Результат використання команди summary()

Наступним чином було створено модель 8.

```
> model_8 <- lm(expenses ~ age + age2 + bmi30*children + bmi + sex + bmi30*smoker + region, data = insurance)
```

Було внесено такі зміни:

– додано нелінійну змінну для віку;

– введено залежність між ожирінням і кількістю дітей;

– введено залежність між ожирінням і курінням.

Як видно на рис. 2.19, значення r-squared дорівнює 0,8664 – це означає, що апроксимація моделі є хорошою, а модель пояснює 86,64% мінливості змінної expenses. Також на рис. 2.18 наведено бета-коефіцієнти.

```
> model_8

Call:
lm(formula = expenses ~ age + age2 + bmi30 * children + bmi + 
    sex + bmi30 * smoker + region, data = insurance)

Coefficients:
    (Intercept)              age             age2            bmi30         children              bmi  
sexmale        smokeryes  
        189.570          -32.151            3.724        -1110.050          625.482          119.763  
-493.646        13407.351  
regionnorthwest  regionsoutheast  regionsouthwest   bmi30:children  bmi30:smokeryes  
       -284.672         -826.108        -1224.315          102.705        19806.915
```

Рисунок 2.18 – Отримані бета-коефіцієнти

```
> summary(model_8)

Call:
lm(formula = expenses ~ age + age2 + bmi30 * children + bmi + 
    sex + bmi30 * smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max 
-17243.5  -1655.5  -1259.4   -719.2  24162.1 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)       189.5704  1367.1462   0.139 0.889739    
age               -32.1505    59.8488  -0.537 0.591223    
age2                3.7240     0.7466   4.988 6.92e-07 ***
bmi30           -1110.0503   477.1893  -2.326 0.020157 *  
children          625.4815   148.8410   4.202 2.82e-05 ***
bmi               119.7631    34.2892   3.493 0.000494 ***
sexmale          -493.6464   244.5170  -2.019 0.043703 *  
smokeryes       13407.3506   440.1157  30.463  < 2e-16 ***
regionnorthwest  -284.6717   349.5485  -0.814 0.415563    
regionsoutheast  -826.1077   351.7673  -2.348 0.018998 *  
regionsouthwest -1224.3152   350.6552  -3.492 0.000496 ***
bmi30:children    102.7048   202.1870   0.508 0.611559    
bmi30:smokeryes 19806.9149   604.8797  32.745  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4446 on 1325 degrees of freedom
Multiple R-squared:  0.8664,    Adjusted R-squared:  0.8652 
F-statistic: 716.1 on 12 and 1325 DF,  p-value: < 2.2e-16
```

Рисунок 2.19 – Результат використання команди summary()

Знаючи, яким чином медичні витрати можуть залежати від характеристик пацієнта, було розроблено останню, модель 9. Було внесено такі зміни:

– додано нелінійну змінну для віку;

– введено залежність між ожирінням і курінням.

Модель було створено так само, як і раніше, із використанням функції lm(), однак було додано створені змінні та ефект взаємодії:

```
> model_final <- lm(expenses ~ age + age2 + children + bmi + sex + bmi30*smoker + region, data = insurance)
```

Як видно на рис. 2.20, значення r-squared дорівнює 0,8664 – це означає, що апроксимація моделі є хорошою, а модель пояснює 86,64% мінливості змінної expenses.

```
> summary(model_final)

Call:
lm(formula = expenses ~ age + age2 + children + bmi + sex + bmi30 *
    smoker + region, data = insurance)

Residuals:
     Min       1Q   Median       3Q      Max
-17297.1  -1656.0  -1262.7   -727.8  24161.6

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       139.0053  1363.1359   0.102 0.918792
age               -32.6181    59.8250  -0.545 0.585690
age2                3.7307     0.7463   4.999 6.54e-07 ***
children          678.6017   105.8855   6.409 2.03e-10 ***
bmi               119.7715    34.2796   3.494 0.000492 ***
sexmale          -496.7690   244.3713  -2.033 0.042267 *
bmi30            -997.9355   422.9607  -2.359 0.018449 *
smokeryes       13404.5952   439.9591  30.468  < 2e-16 ***
regionnorthwest  -279.1661   349.2826  -0.799 0.424285
regionsoutheast  -828.0345   351.6484  -2.355 0.018682 *
regionsouthwest -1222.1619   350.5314  -3.487 0.000505 ***
bmi30:smokeryes 19810.1534   604.6769  32.762  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4445 on 1326 degrees of freedom
Multiple R-squared:  0.8664,    Adjusted R-squared:  0.8653
F-statistic: 781.7 on 11 and 1326 DF,  p-value: < 2.2e-16
```

Рисунок 2.20 – Результат використання команди summary()

2.3 Порівняння моделей

Далі було виконано порівняння якості моделей використовуючи критерії Multiple R-squared та Adjusted R-squared.

Таблиця 2.2 – Порівняння моделей за R-squared

	Multiple R-squared	Adjusted R-squared
Модель 1	0,7509	0,7494
Модель 2	0,7537	0,7521
Модель 3	0,756	0,7541
Модель 4	0,7584	0,7564
Модель 5	0,756	0,7541
Модель 6	0,7583	0,7563
Модель 7	0,8639	0,8628
Модель 8	0,8664	0,8652
Модель 9	0,8664	0,8653

Як можна побачити із таблиці, наведеної вище, моделі з сьомої по дев’яту мають найбільші значення по Multiple R-squared та Adjusted R-squared і тому найбільш точно описують дані. Отже, їх було використано для подальшого прогнозування. Також, усі моделі мають значення p-value, що є меншим за 2.2e-16. Це означає, що предиктори сильно впливають на відповідь [20].

2.4 Висновки до розділу

У даному розділі було проведено аналіз та підготовку даних для створення моделей регресії, які прогнозують медичні витрати на основі характеристик пацієнтів. Було використано набір даних, що містить інформацію про 1338 пацієнтів із США, включаючи вік, стать, індекс маси тіла (ІМТ), кількість дітей, статус куріння та регіон проживання.

Спочатку було здійснено аналіз даних, зокрема перевірено їх розподіл, побудовано матриці кореляції та діаграми розсіювання для виявлення залежностей між змінними. Було визначено, що медичні витрати мають зсув вправо, що свідчить про значну кількість пацієнтів з відносно низькими витратами і деякими з дуже високими витратами.

Далі було побудовано регресійні моделі для прогнозування медичних витрат. Було також враховано нелінійні залежності та взаємодії між змінними, щоб покращити точність прогнозів.

Було створено та порівняно дев'ять моделей. Визначено, що моделі з додаванням нелінійних змінних та взаємодій мають кращі показники R-squared та Adjusted R-squared. Найкращі моделі (7, 8 та 9) пояснюють більше 86% мінливості медичних витрат, що свідчить про задовільну точність прогнозування.

3 ПРОГНОЗУВАННЯ З ВИКОРИСТАННЯМ РЕГРЕСІЙНОЇ МОДЕЛІ

3.1 Застосування моделі до вихідних тренувальних даних

Проаналізувавши оціночні коефіцієнти регресії та статистику відповідності, далі моделі 7, 8 та 9 було використано для прогнозування витрат на медичне страхування. Щоб проілюструвати процес прогнозування, спочатку було застосовано модель до вихідних тренувальних даних, використано функцію predict() наступним чином.

```
> insurance$pred7 <- predict(model_7, insurance)
> insurance$pred8 <- predict(model_8, insurance)
> insurance$pred9 <- predict(model_9, insurance)
```

Таким чином прогнози було збережено як нові вектори з іменем pred8, pred9 та pred9 відповідно у фреймі даних insurance. Далі було обчислено кореляцію між прогнозованою та фактичною вартістю страхування.

```
> cor(insurance$pred7, insurance$expenses)
[1] 0.9294656
> cor(insurance$pred8, insurance$expenses)
[1] 0.9308139
> cor(insurance$pred9, insurance$expenses)
[1] 0.9307999
```

Кореляція 0,93 вказує на дуже сильну лінійну залежність між прогнозованими та фактичними значеннями. Це означає, що моделі є достатньо точними.

Також було представлено результати у вигляді діаграм розсіювання (рис 3.1-3.3). За допомогою наступних R-команд було побудовано відношення та додано одиничну пряму – лінію зі зсувом, рівним 0, та нахилом, рівним 1. Параметри col, lwd та lty визначають колір, товщину та тип лінії відповідно.

```
> plot(insurance$pred7, insurance$expenses)
> abline(a = 0, b = 1, col = "blue", lwd = 3, lty = 2)
```

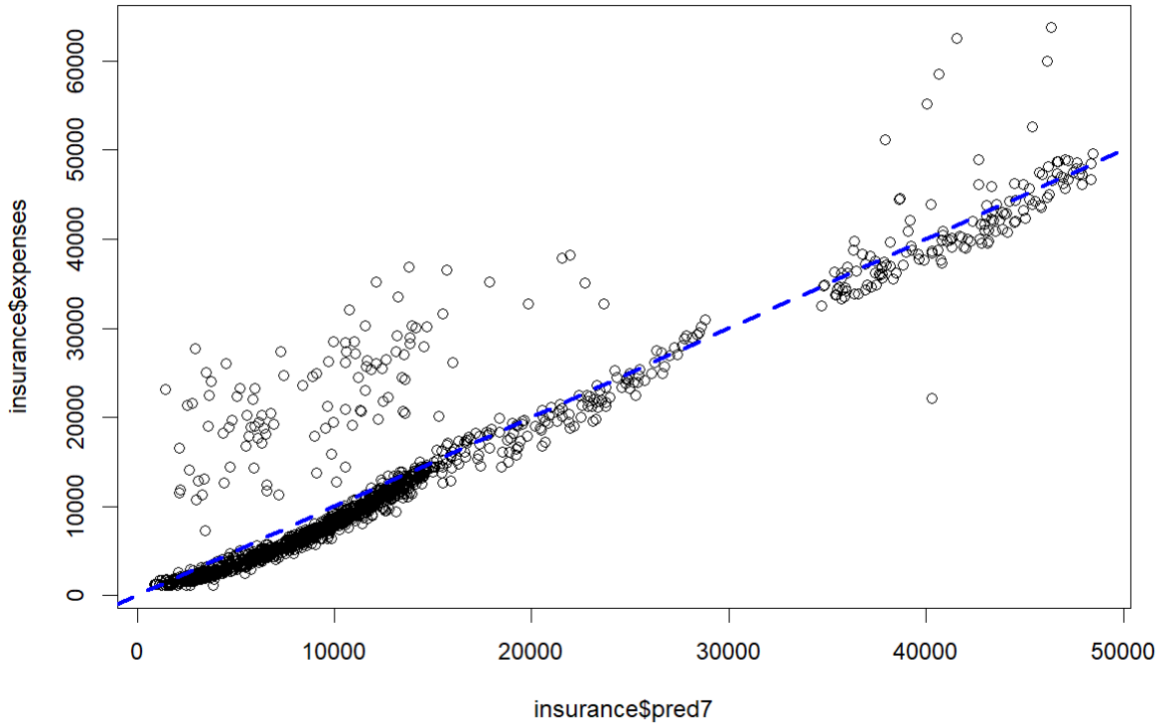

Рисунок 3.1 – Діаграма розсіювання прогнозу на основі моделі 7

```
> plot(insurance$pred8, insurance$expenses)
> abline(a = 0, b = 1, col = "blue", lwd = 3, lty = 2)
```

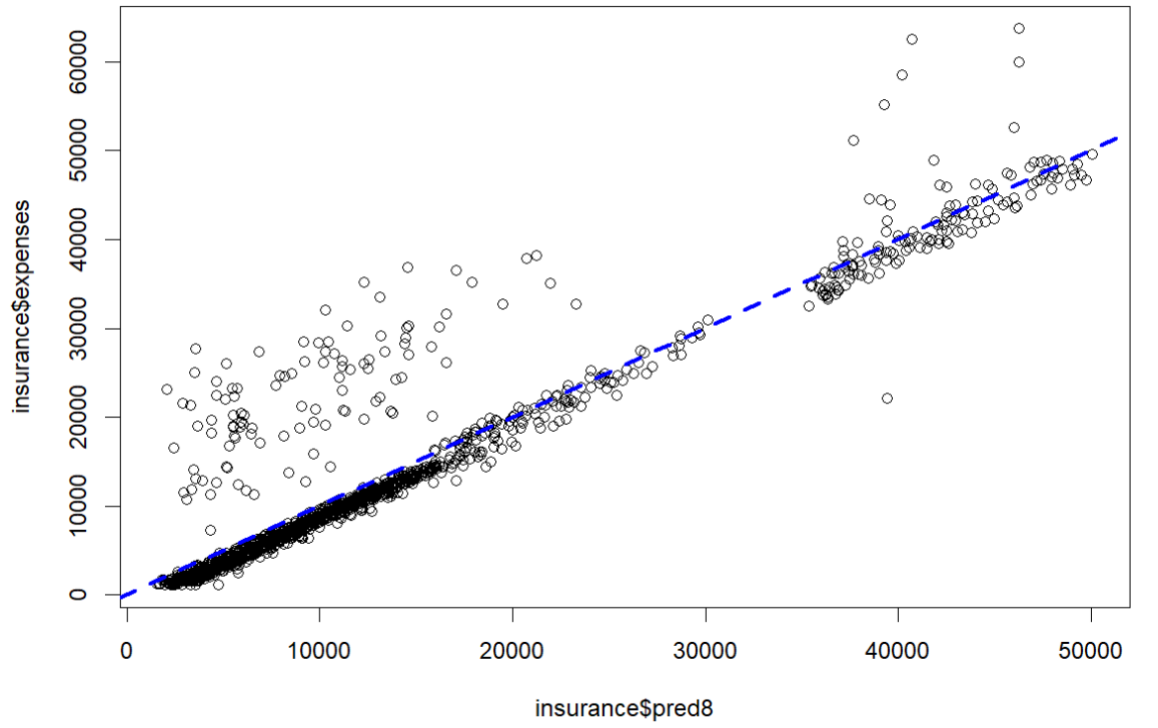

Рисунок 3.2 – Діаграма розсіювання прогнозу на основі моделі 8

```
> plot(insurance$pred9, insurance$expenses)
> abline(a = 0, b = 1, col = "blue", lwd = 3, lty = 2)
```

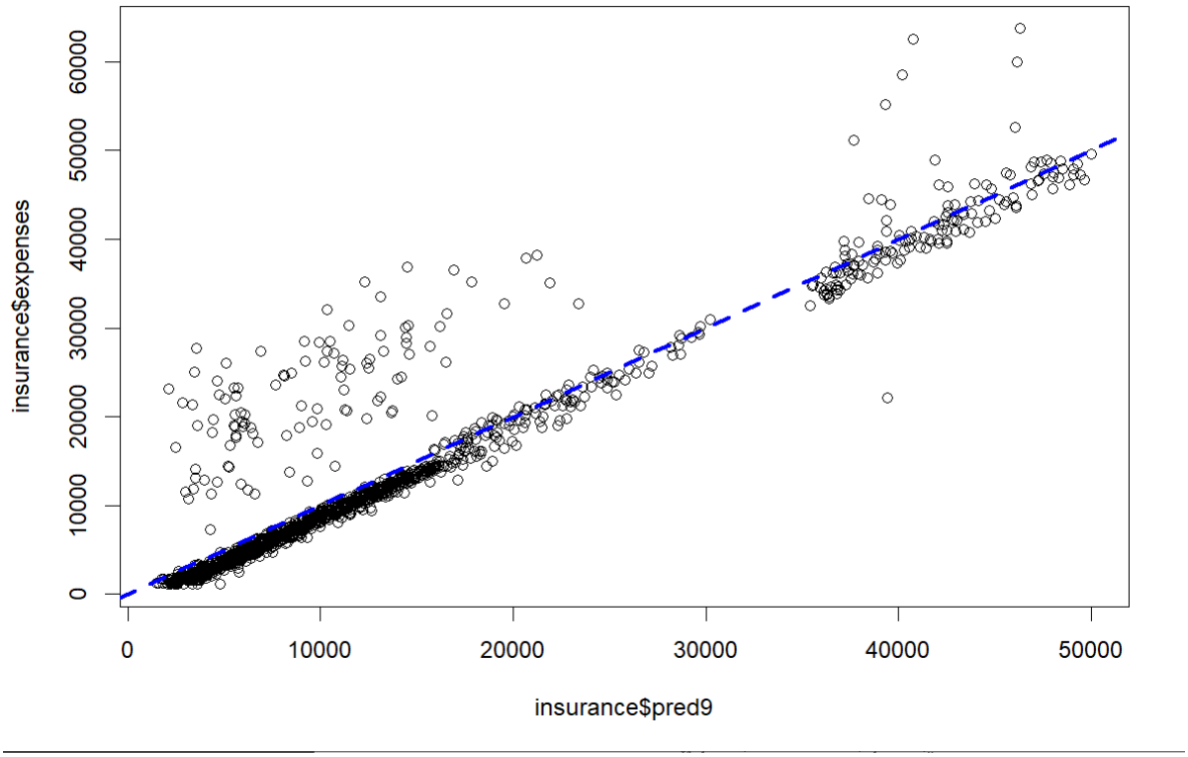

Рисунок 3.3 – Діаграма розсіювання прогнозу на основі моделі 9

Точки, що потрапляють на діагональну пунктирну лінію y = x або розташовані поруч з нею, вказують на прогнози, дуже близькі до фактичних значень.

Точки, розташовані вище діагональної прямої, відповідають випадкам, коли фактичні витрати виявилися більшими, ніж очікувалося, а точки, розташовані нижче цієї лінії, – випадкам, коли витрати були меншими, ніж очікувалося. Тут можна побачити, що невелика кількість пацієнтів із набагато більшими, ніж очікувалося, медичними витратами врівноважується великою кількістю пацієнтів із витратами, трохи меншими за очікувані.

Далі для порівняння прогнозів було використано середню абсолютну похибку у відсотках (MAPE). Вона часто використовується як функція втрат для задач регресії та при оцінюванні моделей, оскільки вона має дуже інтуїтивно

зрозуміле тлумачення з точки зору відносної похибки [21,22]. Формула має наступний вигляд.

$$MAPE = \frac{1}{N}\sum_{j=1}^{N}\frac{|y_j - \hat{y}_j|}{y_j} \times 100, \text{ де}$$

y – фактична оцінка, $\hat{y}$ – прогнозована, оцінка, N – кількість прогнозованих оцінок.

З метою визначення MAPE для моделі 7 було виконано наступний код.

```
     > insurance$pred7_i <- predict(model_7, insurance, interval="predict",
level = .95)
     > compare_model7 <- cbind (insurance$expenses, insurance$pred7_i[,1])
     > colnames(compare_model7) <- c("actual", "predicted")
     > summary(compare_model7)
          actual        predicted
      Min.   : 1122   Min.   :  888.9
      1st Qu.: 4740   1st Qu.: 6063.8
      Median : 9382   Median :10293.7
      Mean   :13270   Mean   :13270.4
      3rd Qu.:16640   3rd Qu.:14034.4
      Max.   :63770   Max.   :48469.3
     >         mape_model_7         <-         (sum(abs(compare_model7[,1]-
compare_model7[,2])/abs(compare_model7[,1]))/nrow(compare_model7))*100
     > mape_model_7
     [1] 25.758
```

Як можна побачити, для моделі 7 значення критерію MAPE дорівнює 25,758, що означає похибку у 25,758%. Аналогічний код було виконано для визначення MAPE моделі 8.

```
     > insurance$pred8_i <- predict(model_8, insurance, interval="predict",
level = .95)
     > compare_model8 <- cbind (insurance$expenses, insurance$pred8_i[,1])
     > colnames(compare_model8) <- c("actual", "predicted")
     > summary(compare_model8)
          actual        predicted
      Min.   : 1122   Min.   : 1576
      1st Qu.: 4740   1st Qu.: 5803
      Median : 9382   Median : 9756
```

```
      Mean   :13270   Mean   :13270
      3rd Qu.:16640   3rd Qu.:14785
      Max.   :63770   Max.   :50057
     >         mape_model_8         <-         (sum(abs(compare_model8[,1]-
compare_model8[,2])/abs(compare_model8[,1]))/nrow(compare_model8))*100
     > mape_model_8
     [1] 27.36326
```

Як можна побачити, для моделі 8 значення критерію MAPE дорівнює 27,36326, що означає похибку у 27,36326%. Аналогічний код було виконано для визначення MAPE моделі 9.

```
     > insurance$pred9_i <- predict(model_9, insurance, interval="predict",
level = .95)
     > compare_model9 <- cbind (insurance$expenses, insurance$pred9_i[,1])
     > colnames(compare_model9) <- c("actual", "predicted")
     > summary(compare_model9)
          actual        predicted
      Min.   : 1122   Min.   : 1519
      1st Qu.: 4740   1st Qu.: 5822
      Median : 9382   Median : 9782
      Mean   :13270   Mean   :13270
      3rd Qu.:16640   3rd Qu.:14803
      Max.   :63770   Max.   :50013
     >         mape_model_9         <-         (sum(abs(compare_model9[,1]-
compare_model9[,2])/abs(compare_model9[,1]))/nrow(compare_model9))*100
     > mape_model_9
     [1] 27.35791
```

Як можна побачити, для моделі 9 значення критерію MAPE дорівнює 27,35791, що означає похибку у 27,35791%. Після цього було створено порівняльну таблицю.

Наступним кроком було оцінювання моделей за критерієм MSE. Цей критерій оцінює якість або передбачувача (тобто функції, що відображує довільні входи до вибірки значень деякої випадкової величини), або оцінювача (тобто математичної функції, що відображує вибірку даних до оцінки параметра сукупності, з якої відбираються ці дані) [23].

Для моделі 7 було виконано наступний код.

```
> mse_model_7 <- sum(mean((compare_model7[,1] - compare_model7[,2])^2))/nrow(compare_model7)
> mse_model_7
[1] 14905.48
```

Для моделі 8 було виконано наступний код.

```
> mse_model_8 <- sum(mean((compare_model8[,1] - compare_model8[,2])^2))/nrow(compare_model8)
> mse_model_8
[1] 14630.79
```

Для моделі 9 було виконано наступний код.

```
> mse_model_9 <- sum(mean((compare_model9[,1] - compare_model9[,2])^2))/nrow(compare_model9)
> mse_model_9
[1] 14633.64
```

Наступним кроком було оцінювання моделей за критерієм RMSE. Це квадратичне середнє відхилення вибірки. По суті є квадратичним середнім різниць між спостережуваними значеннями та прогнозованими [24].

Для моделі 7 було виконано наступний код.

```
> rmse_model_7 <- sum(sqrt(mean((compare_model7[,1] - compare_model7[,2])^2)))/nrow(compare_model7)
> rmse_model_7
[1] 3.337682
```

Для моделі 8 було виконано наступний код.

```
> rmse_model_8 <- sum(sqrt(mean((compare_model8[,1] - compare_model8[,2])^2)))/nrow(compare_model8)
> rmse_model_8
[1] 3.306784
```

Для моделі 9 було виконано наступний код.

```
> rmse_model_9 <- sum(sqrt(mean((compare_model9[,1] - compare_model9[,2])^2)))/nrow(compare_model9)
> rmse_model_9
[1] 3.307106
```

Наступним кроком було оцінювання моделей за критерієм MRE. Цей критерій вимірює середню відносну похибку, тобто наскільки прогнози моделі відрізняються у відсотках від фактичних значень [25].

Для моделі 7 було виконано наступний код.

```
> mre_model_7 <- sum(abs(compare_model7[,2]-compare_model7[,1]/compare_model7[,1]))/nrow(compare_model7)
> mre_model_7
[1] 13269.42
```

Для моделі 8 було виконано наступний код.

```
> mre_model_8 <- sum(abs(compare_model8[,2]-compare_model8[,1]/compare_model8[,1]))/nrow(compare_model8)
> mre_model_8
[1] 13269.42
```

Для моделі 9 було виконано наступний код.

```
> mre_model_9 <- sum(abs(compare_model9[,2]-compare_model9[,1]/compare_model9[,1]))/nrow(compare_model9)
> mre_model_9
[1] 13269.42
```

Таблиця 2.2 – Порівняння моделей за критерієм MAPE, MSE, RMSE, MRE

	MAPE	MSE	RMSE	MRE
Модель 7	25,758%	14905,48	3,337682	13269,42
Модель 8	27,36326%	14630,79	3,306784	13269,42
Модель 9	27,35791%	14633,64	3,307106	13269,42

Отже, було обрано модель 9 для прогнозування нових клієнтів.

3.2 Прогнозування для нових учасників

Далі було виконано прогнозування витрат для нових учасників програми страхування використовуючи модель 9. Для функції predict() було надано фрейм даних з інформацією про передбачуваних пацієнтів. Наприклад, щоб оцінити

витрати на страхування 30-річного чоловіка, який не курить, але має надмірну вагу, проживає з двома дітьми на північному сході США, було використано наступний код:

```
> predict(model_9, data.frame(age = 30, age2 = 30^2, children = 2, bmi = 30, sex = "male", bmi30 = 1, smoker = "no", region = "northeast"))
5973.774
```

Використовуючи це значення, страхова компанія, можливо, вирішить встановити ціну для даної демографічної групи на рівні близько 6000 доларів на рік, або 500 доларів на місяць, щоб гарантувати беззбитковість. Щоб порівняти цей показник для жінки з такими ж характеристиками, було використано функцію predict() схожим чином:

```
> predict(model_9, data.frame(age = 30, age2 = 30^2, children = 2, bmi = 30, sex = "female", bmi30 = 1, smoker = "no", region = "northeast"))
6470.543
```

Можна побачити, що різниця між цими двома значеннями складає 496,769 і дорівнює розрахунковому коефіцієнту регресійної моделі для sexmale. За оцінками, в середньому чоловіки за інших рівних умов витрачають на медицину приблизно на 496 доларів на рік менше, ніж жінки.

Це свідчить про те, що прогнозовані витрати є сумою кожного з коефіцієнтів регресії, помноженого на відповідне значення у фреймі даних прогнозування. Наприклад, використовуючи коефіцієнт регресії моделі 678,6017 для кількості дітей, можна припустити, що ті, у кого немає дітей, витрачають на медичне обслуговування на 1357,203 долара менше, що і було підтверджено:

```
> predict(model_9, data.frame(age = 30, age2 = 30^2, children = 0, bmi = 30, sex = "female", bmi30 = 1, smoker = "no", region = "northeast"))
5113.34
```

Виконавши аналогічні операції ще для кількох клієнтських сегментів, страхова компанія може розробити вигідну структуру ціноутворення для різних демографічних показників. Експорт коефіцієнтів регресійної моделі дозволяє побудувати власну функцію прогнозування. Одним з можливих варіантів

використання такої функції було б впровадження регресійної моделі в базу даних клієнтів для прогнозування в реальному часі.

3.3 Висновки до розділу

У даному розділі було застосовано побудовані регресійні моделі до вихідних тренувальних даних та здійснено прогнозування витрат на медичне страхування. Моделі 7, 8 та 9 було використано для прогнозування медичних витрат, було продемонстровано задовільну точність прогнозів із кореляцією між фактичними та прогнозованими витратами на рівні приблизно 0,93.

Засновуючись на діаграмах розсіювання, можна сказати, що більшість точок розташовані поблизу діагональної лінії, що також вказує на задовільну точність моделей. Після порівняння моделей за різними критеріями, такими як MAPE, MSE, RMSE та MRE, було визначено, що всі три моделі мають подібні показники ефективності, але модель 9 є більш прийнятною.

Також було виконано прогнозування для нових учасників програми страхування використовуючи обрану модель 9, що продемонструвало практичне застосування моделей для реальних клієнтів. Моделі може бути використано для оцінки витрат на медичне страхування для різних демографічних груп із достатньою точністю, що дозволить страховим компаніям розробити вигідну структуру ціноутворення.

ВИСНОВКИ

Протягом виконання даної роботи було розглянуто процес аналізу та прогнозування медичних витрат з використанням регресійних моделей на основі характеристик пацієнтів. Дослідження складалося з кількох ключових етапів.

Складовими першого етапу є аналіз та підготовка даних. Було використано набір даних, що містив медичні витрати для 1338 пацієнтів із США. Було проведено детальний аналіз даних, включаючи перевірку розподілу змінних, побудову матриць кореляції та діаграм розсіювання. Визначено, що медичні витрати мають зсув вправо, що свідчить про значну кількість пацієнтів з відносно низькими витратами та деяких з дуже високими витратами.

На другому етапі було створено регресійні моделі для прогнозування медичних витрат. Враховувалися такі змінні, як вік, індекс маси тіла (ІМТ), кількість дітей, стать, статус куріння та регіон проживання. Було додано нелінійні залежності та взаємодії між змінними для покращення точності прогнозів. Після порівняння дев'яти моделей було визначено, що моделі з додаванням нелінійних змінних та взаємодій мають кращі показники R-squared та Adjusted R-squared, пояснюючи більше 86% мінливості медичних витрат.

На третьому етапі моделі було застосовано до вихідних тренувальних даних для прогнозування медичних витрат. Після порівняння прогнозів з фактичними значеннями було визначено задовільну точність моделей із кореляцією між фактичними та прогнозованими витратами на рівні приблизно 0,93. Для оцінювання моделей було використано середню абсолютну похибку у відсотках (MAPE), середньоквадратичну похибку (MSE), квадратичне середнє відхилення (RMSE) та середню відносну похибку (MRE). Далі було обрано кращу модель.

Останнім етапом було проведено прогнозування витрат для нових учасників програми страхування на основі кращої моделі. В результаті було визначено, що моделі можуть бути використані для прогнозування витрат на медичне страхування для різних демографічних груп із задовільною точністю.

ПЕРЕІЛК ДЖЕРЕЛ ПОСИЛАННЯ

1. Інформаційна система – Вікіпедія. URL: https://uk.wikipedia.org/wiki/%D0%86%D0%BD%D1%84%D0%BE%D1%80%D0%BC%D0%B0%D1%86%D1%96%D0%B9%D0%BD%D0%B0_%D1%81%D0%B8%D1%81%D1%82%D0%B5%D0%BC%D0%B0 (дата звернення: 29.05.2024).

2. Information Systems: Definitions and Components. URL: https://www.uotechnology.edu.iq/ce/Lectures/SarmadFuad-MIS/MIS_Lecture_3.pdf (дата звернення: 29.05.2024).

3. Державна служба статистики України – Населення та соціальна статистика. URL: https://stat.gov.ua/uk/topics/naselennya-ta-sotsialna-statystyka (дата звернення: 05.05.2024).

4. Forinsurer – Підсумки страхового ринку України. URL: https://forinsurer.com/news/22/05/30/41284 (дата звернення: 05.05.2024).

5. Інформаційне забезпечення інноваційної діяльності страхових компаній. URL: https://kerivnyk.info/2013/01/morgun.html (дата звернення: 29.05.2024).

6. R (мова програмування) – Вікіпедія. URL: https://uk.wikipedia.org/wiki/R_(%D0%BC%D0%BE%D0%B2%D0%B0_%D0%BF%D1%80%D0%BE%D0%B3%D1%80%D0%B0%D0%BC%D1%83%D0%B2%D0%B0%D0%BD%D0%BD%D1%8F) (дата звернення: 05.05.2024).

7. RStudio – Вікіпедія. URL: https://uk.wikipedia.org/wiki/RStudio (дата звернення: 29.05.2024).

8. Working with CSV files in R Programming - GeeksforGeeks. URL: https://www.geeksforgeeks.org/working-with-csv-files-in-r-programming/ (дата звернення: 29.05.2024).

9. Кореляція – Вікіпедія. URL: https://uk.wikipedia.org/wiki/%D0%9A%D0%BE%D1%80%D0%B5%D0%BB%D1%8F%D1%86%D1%96%D1%8F (дата звернення: 29.05.2024).

10. Точкова діаграма – Вікіпедія. URL: https://uk.wikipedia.org/wiki/%D0%A2%D0%BE%D1%87%D0%BA%D0%BE%D0%B2%D0%B0_%D0%B4%D1%96%D0%B0%D0%B3%D1%80%D0%B0%D0%BC%D0%B0 (дата звернення: 29.05.2024).

11. Did You Know? | LOESS | National Centers for Environmental Information (NCEI). URL: https://www.ncei.noaa.gov/access/monitoring/dyk/loess (дата звернення: 29.05.2024).

12. lm function - RDocumentation. URL: https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/lm (дата звернення: 29.05.2024).

13. Beta coefficients in linear models. Statistics for Ecologists Exercises. URL: https://www.dataanalytics.org.uk/beta-coefficients-from-linear-models/ (дата звернення: 29.05.2024).

14. Using Categorical Data with One Hot Encoding. URL: https://www.kaggle.com/code/dansbecker/using-categorical-data-with-one-hot-encoding (дата звернення: 29.05.2024).

15. summary function – RDocumentation. URL: https://www.rdocumentation.org/packages/base/versions/3.6.2/topics/summary (дата звернення: 29.05.2024).

16. Regression Techniques in Machine Learning. URL: https://www.analyticsvidhya.com/blog/2015/08/comprehensive-guide-regression/ (дата звернення: 29.05.2024).

17. What Is Linear Regression? | IBM. URL: https://www.ibm.com/topics/linear-regression/ (дата звернення: 29.05.2024).

18. Consequences of chronic diseases and other limitations associated with old age – a scoping review | BMC Public Health | Full Text. URL: https://bmcpublichealth.biomedcentral.com/articles/10.1186/s12889-019-7762-5 (дата звернення: 29.05.2024).

19. R-Squared: Definition, Calculation Formula, Uses, and Limitations. URL: https://www.investopedia.com/terms/r/r-squared.asp (дата звернення: 29.05.2024).

20. P-Value in Statistical Hypothesis Tests: What is it? - Statistics How To. URL: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/p-value/ (дата звернення: 29.05.2024).

21. Mean absolute percentage error. URL: https://en.wikipedia.org/wiki/Mean_absolute_percentage_error (дата звернення: 29.05.2024).

22. How to Calculate Mean Absolute Percentage Error (MAPE) in R | R-bloggers. URL: https://www.r-bloggers.com/2021/08/how-to-calculate-mean-absolute-percentage-error-mape-in-r/ (дата звернення: 29.05.2024).

23. Середньоквадратична похибка – Вікіпедія. URL: https://uk.wikipedia.org/wiki/%D0%A1%D0%B5%D1%80%D0%B5%D0%B4%D0%BD%D1%8C%D0%BE%D0%BA%D0%B2%D0%B0%D0%B4%D1%80%D0%B0%D1%82%D0%B8%D1%87%D0%BD%D0%B0_%D0%BF%D0%BE%D1%85%D0%B8%D0%B1%D0%BA%D0%B0 (дата звернення: 29.05.2024).

24. Root mean square deviation. URL: https://en.wikipedia.org/wiki/Root_mean_square_deviation (дата звернення: 29.05.2024).

25. MeanRelativeError. URL: https://haibal.com/documentation/metric-mean-relative-error/ (дата звернення: 29.05.2024).

ДОДАТОК А

Лістинг коду

```
# # Прогнозування медичних витрат
# Дослідження та підготовка даних
insurance <- read.csv("insurance.csv", stringsAsFactors = TRUE)
str(insurance)

# summarize для змінної витрат
summary(insurance$expenses)

# гістограма витрат на страхування
hist(insurance$expenses)

# таблиця регіону
table(insurance$region)

# дослідження взаємозв'язків між ознаками: матриця кореляцій
cor(insurance[c("age", "bmi", "children", "expenses")])

# візуалізація взаємозв'язків між ознаками: матриця точкових діаграм
pairs(insurance[c("age", "bmi", "children", "expenses")])

# більш інформативна матриця точкових діаграм
library(psych)
pairs.panels(insurance[c("age", "bmi", "children", "expenses")])

# -Навчання
# Створення моделей
model_1 <- lm(expenses ~ age + children + bmi + sex + smoker + region,
              data = insurance)

# перегляд бета-коефіцієнтів
model_1

# Оцінка продуктивності моделі
# перегляд більш детальної інформації про бета-коефіцієнти
summary(model_1)


# Покращення ефективності моделі
# додано вищий порядок для змінної "вік"
insurance$age2 <- insurance$age^2

model_2 <- lm(expenses ~ age + age2 + children + bmi + sex + smoker + region,
              data = insurance)

model_2
summary(model_2)

# додано індикатор для ІМТ >= 30
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)

model_3 <- lm(expenses ~ age + children + bmi30*bmi + sex + smoker + region,
              data = insurance)

model_3
summary(model_3)
```

```
model_4 <- lm(expenses ~ age + age2 + children + bmi30*bmi + sex + smoker + region,
              data = insurance)

model_4
summary(model_4)

model_5 <- lm(expenses ~ age + bmi30*children + bmi + sex + smoker + region,
              data = insurance)

model_5
summary(model_5)

model_6 <- lm(expenses ~ age + age2 + bmi30*children + bmi + sex + smoker + region,
              data = insurance)

model_6
summary(model_6)

model_7 <- lm(expenses ~ age + bmi30*children + bmi + sex + bmi30*smoker + region,
              data = insurance)

model_7
summary(model_7)

model_8 <- lm(expenses ~ age + age2 + bmi30*children + bmi + sex + bmi30*smoker + region,
              data = insurance)

model_8
summary(model_8)

# створення фінальної моделі
model_9 <- lm(expenses ~ age + age2 + children + bmi + sex +
                bmi30*smoker + region, data = insurance)

summary(model_9)


# прогнозування за допомогою покращенної регресійної моделі
# застосування моделі до вихідних тренувальних даних
insurance$pred7 <- predict(model_7, insurance)
cor(insurance$pred7, insurance$expenses)

insurance$pred8 <- predict(model_8, insurance)
cor(insurance$pred8, insurance$expenses)

insurance$pred9 <- predict(model_9, insurance)
cor(insurance$pred9, insurance$expenses)

insurance$pred7_i <- predict(model_7, insurance, interval="predict", level = .95)
compare_model7 <- cbind (insurance$expenses, insurance$pred7_i[,1])
colnames(compare_model7) <- c("actual", "predicted")
summary(compare_model7)
mape_model_7 <- (sum(abs(compare_model7[,1]-
compare_model7[,2])/abs(compare_model7[,1]))/nrow(compare_model7))*100
mape_model_7

mse_model_7 <- sum(mean((compare_model7[,1] - compare_model7[,2])^2))/nrow(compare_model7)
mse_model_7
```

```
rmse_model_7 <- sum(sqrt(mean((compare_model7[,1] - compare_model7[,2])^2)))/nrow(compare_model7)
rmse_model_7

mre_model_7 <- sum(abs(compare_model7[,2]-
compare_model7[,1]/compare_model7[,1]))/nrow(compare_model7)
mre_model_7

insurance$pred8_i <- predict(model_8, insurance, interval="predict", level = .95)
compare_model8 <- cbind (insurance$expenses, insurance$pred8_i[,1])
colnames(compare_model8) <- c("actual", "predicted")
summary(compare_model8)
mape_model_8 <- (sum(abs(compare_model8[,1]-
compare_model8[,2])/abs(compare_model8[,1]))/nrow(compare_model8))*100
mape_model_8

mse_model_8 <- sum(mean((compare_model8[,1] - compare_model8[,2])^2))/nrow(compare_model8)
mse_model_8

rmse_model_8 <- sum(sqrt(mean((compare_model8[,1] - compare_model8[,2])^2)))/nrow(compare_model8)
rmse_model_8

mre_model_8 <- sum(abs(compare_model8[,2]-
compare_model8[,1]/compare_model8[,1]))/nrow(compare_model8)
mre_model_8

insurance$pred9_i <- predict(model_9, insurance, interval="predict", level = .95)
compare_model9 <- cbind (insurance$expenses, insurance$pred9_i[,1])
colnames(compare_model9) <- c("actual", "predicted")
summary(compare_model9)
mape_model_9 <- (sum(abs(compare_model9[,1]-
compare_model9[,2])/abs(compare_model9[,1]))/nrow(compare_model9))*100
mape_model_9

mse_model_9 <- sum(mean((compare_model9[,1] - compare_model9[,2])^2))/nrow(compare_model9)
mse_model_9

rmse_model_9 <- sum(sqrt(mean((compare_model9[,1] - compare_model9[,2])^2)))/nrow(compare_model9)
rmse_model_9

mre_model_9 <- sum(abs(compare_model9[,2]-
compare_model9[,1]/compare_model9[,1]))/nrow(compare_model9)
mre_model_9


# діаграма розсіювання
plot(insurance$pred7, insurance$expenses)
abline(a = 0, b = 1, col = "blue", lwd = 3, lty = 2)

plot(insurance$pred8, insurance$expenses)
abline(a = 0, b = 1, col = "blue", lwd = 3, lty = 2)

plot(insurance$pred9, insurance$expenses)
abline(a = 0, b = 1, col = "blue", lwd = 3, lty = 2)

#Для нових клієнтів
predict(model_9,
      data.frame(age = 30, age2 = 30^2, children = 2,
            bmi = 30, sex = "male", bmi30 = 1,
            smoker = "no", region = "northeast"))
```

```
predict(model_9,
      data.frame(age = 30, age2 = 30^2, children = 2,
            bmi = 30, sex = "female", bmi30 = 1,
            smoker = "no", region = "northeast"))

predict(model_9,
      data.frame(age = 30, age2 = 30^2, children = 0,
            bmi = 30, sex = "female", bmi30 = 1,
            smoker = "no", region = "northeast"))
```