Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Центр Післядипломної освіти
(повна назва)
Кафедра Штучного інтелекту
(повна назва)

КВАЛІФІКАЦІЙНА РОБОТА
Пояснювальна записка

рівень вищої освіти другий (магістерський)

Класифікація страхових випадків методами машинного навчання

(тема)

Виконав:
студент 2 курсу, групи СШІмзд-22-1
Зубова В.В.
(прізвище, ініціали)

Спеціальність 122 Комп'ютерні науки
(код і повна назва спеціальності)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва спеціалізації)
Керівник проф. Аврунін О.Г.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2024 р.

Харківський національний університет радіоелектроніки

Центр Післядипломної освіти
(повна назва)
Кафедра Штучного інтелекту
(повна назва)
Рівень вищої освіти другий (магістерський)

Спеціальність 122 Комп'ютерні науки
(код і повна назва)
Тип програми освітньо-наукова
(освітньо-професійна або освітньо-наукова)
Освітня програма Системи штучного інтелекту
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«_____« _____ 20 ___ р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

студентові Зубовій Віталіні Вікторівні
(прізвище, ім'я, по батькові)

1. Тема роботи Класифікація страхових випадків методами машинного навчання

затверджена наказом університету від 22 квітня 20 24 р. № 61Стз
2. Термін подання студентом роботи до екзаменаційної комісії 13 червня 20 24 р.
3. Вихідні дані до роботи Науково-технічні публікації, дані Інтернет-джерел щодо методів машинного навчання

4. Перелік питань, що потрібно опрацювати в роботі
1) Огляд предметної галузі
2) Аналіз бази страхових випадків та попередня обробка даних
3) Побудова та тестування моделей класифікації страхових випадків

КАЛЕНДАРНИЙ ПЛАН

№	Назва етапів роботи	Терміни виконання етапів роботи	Примітка
1	Отримання завдання на кваліфікаційну роботу	22.04.2024	виконано
2	Аналіз бази страхових випадків	27.04.2024	виконано
3	Попередня обробка даних	04.05.2024	виконано
4	Вирішення проблеми незбалансованої вибірки	11.05.2024	виконано
5	Побудова базової моделі класифікації страхових	18.05.2024	виконано
	випадків		
6	Побудова та тестування класифікаторів	25.06.2024	виконано
7	Порівняння класифікаторів страхових випадків	01.06.2024	виконано
8	Написання пояснювальної записки	08.06.2024	виконано
9	Захист перед ЕК	13.06.2024	

Дата видачі завдання 22 квітня 20 24 р.

Студент _____
(підпис)

Керівник роботи _____ проф. Аврунін О.Г.
(підпис) (посада, прізвище, ініціали)

РЕФЕРАТ

Пояснювальна записка: 119 с., 45 рис., 29 табл., 3 дод., 33 джерела.

КЛАСИФІКАЦІЯ, МАШИННЕ НАВЧАННЯ, НЕЗБАЛАНСОВАНА ВИБІРКА, РИНОК, СТРАХОВИЙ ВИПАДОК, СТРАХУВАННЯ, ШАХРАЙСЬКІ ПРЕТЕНЗІЇ.

Об'єктом дослідження є база даних страхових випадків, що містить інформацію про встановлений чи відсутній факт шахрайства за клієнтською претензією, а також закономірності утворення, існування та властивості, таких випадків.

Предметом дослідження виступає сукупність механізмів побудови та використання аналітичних та прогнозних моделей щодо страхових випадків.

Мета дослідження полягає у проведенні аналізу бази даних страхових претензій, створенні на його основі прогнозних моделей класифікації, виявленні закономірностей існування страхових випадків та факторів, що визначають принципи їх утворення, використовуючи методи машинного навчання та інтелектуального аналізу даних, для подальшого попередження шахрайських дій та запобігання можливих небажаних збитків представників страхового бізнесу.

Методи дослідження: кореляційний та статистичний аналізи, методи машинного навчання: логістична регресія, метод опорних векторів, алгоритм k-найближчих сусідів, Байєсовський класифікатор, дерево рішень, випадковий ліс та нейронна мережа.

ABSTRACT

Master’s thesis contains: 119 pp., 45 fig., 29 tabl., 3 ann., 33 references.

CLASSIFICATION, FRAUDULENT CLAIMS, INSURANCE CASE, INSURANCE MARKET, MACHINE LEARNING, UNBALANCED SAMPLE.

The object of the study is a database of insurance cases, which contains information about established or absent fact of fraud on the client's claim, as well as patterns of formation, existence and properties of such cases.

The subject of the study is a set of mechanisms for building and using analytical and predictive models for insurance cases.

The purpose of the research is to conduct an analysis of the database of insurance claims, to create predictive classification models based on it, to identify patterns in the existence of insurance cases and factors that determine the principles of their formation, using methods of machine learning and intelligent data analysis, for further prevention of fraudulent actions and possible prevention unwanted losses of insurance business representatives.

Research methods: correlation and statistical analysis, machine learning methods: logistic regression, support vector method, k-nearest neighbor algorithm, Bayesian classifier, decision tree, random forest and neural network.

ЗМІСТ

Вступ 8
1 Огляд предметної галузі 10
1.1 Аналіз ринку страхування України 10
1.2 Застосування методів машинного навчання у бізнес-цілях 20
1.3 Вирішення завдань страхування методами машинного навчання 24
1.4 Висновки до розділу 1 31
2 Аналіз бази страхових випадків та попередня обробка даних 33
2.1 Постановка та обґрунтування задач роботи 33
2.2 Підготовка даних до моделювання 35
2.3 Візуалізація та оцінка статистичних характеристик бази 45
2.4 Висновки до розділу 2 53
3 Побудова та тестування моделей класифікації страхових випадків 55
3.1 Вирішення проблеми незбалансованої вибірки 55
3.2 Побудова базової моделі 60
3.2.1 Логістична регресія на усіх змінних 60
3.2.2 Логістична регресія на значущих змінних 65
3.3 Побудова нелінійних класифікаторів 67
3.3.1 Метод опорних векторів 67
3.3.2 Метод k-найближчих сусідів 69
3.3.3 Баєсова класифікація 72
3.3.4 Дерево рішень 74
3.3.5 Випадковий ліс 75
3.3.6 Класифікаційна нейронна мережа (CNN) 77
3.4 Висновки до розділу 3 79
Висновки 83
Перелік джерел посилання 86
Додаток А Опис змінних датасету Kaggle 90

Додаток Б Фрагмент виконання коду з моделювання класифікаторів на даних вибірки 2 SMOTE 91
Додаток В Відомість кваліфікаційної роботи..... 119

ВСТУП

Актуальність теми. Одним з найважливіших фінансових показників для страхових організацій є різниця між вартістю проданих страховок і витратами на відшкодування за страховими випадками. Однак розвинуті страхові ринки потерпають від великої кількості шахрайських випадків, які є причиною утворення значної суми збитків. Одним із найпопулярніших видів страхування, де спостерігається вагома частка шахрайських дій, є автострахування. Скорочення витрат є критично важливою метою, для реалізації якої страховики використовують різноманітні методи. Традиційні підходи аналізу таких випадків можуть бути мало ефективними та не здатні виявити приховані залежності. На відміну від експертних, методи машинного навчання засновуються на аналізі великої кількості показників, що мають суттєвий вплив, виявленні серед них неявних закономірностей та допомагають врахувати більшу частину ознак, що пояснюють моделі поведінки споживачів страхових послуг чи слугують для побудови прогнозних моделей задля мінімізації можливих ризиків. Запровадження машинного навчання дозволить отримати більш високу точність прогнозів завдяки урахуванню збільшеного числа факторів, що впливають на ймовірність страхового випадку та можливості проведення дослідження на великому об’ємі даних, що були накопичені за великий проміжок часу. Алгоритми машинного навчання можуть виявляти закономірності, які здаються не пов'язаними один з одним або залишаються непоміченими людиною.

Тематика виявлення шахрайських претензій у автострахуванні є доволі поширеною серед зарубіжних дослідників, у той час коли для робіт вітчизняних науковців дана проблема не є популярною. Тому я вважаю актуальним проведення дослідження даних зі страхових випадків задля виявлення факторів, які формують закономірності утворення шахрайства та побудови на їх основі моделей класифікації.

В кваліфікаційній роботі поставлені та розв’язуються наступні завдання:

– аналіз предметної області: аналіз ринку страхування України;

– вивчення теоретичних засад використання машинного навчання у страховій галузі;

– пошук, огляд і аналіз публікацій та досліджень щодо застосування методів машинного навчання в страховому бізнесі;

– постановка та обґрунтування задач роботи;

– первинний аналіз бази даних претензій з автострахування в рамках задачі аналізу страхових випадків;

– підготовка даних до стану, придатного до моделювання;

– оцінка статистичних характеристик датасета для аналізу страхових випадків;

– дослідження найвпливовіших факторів, проведення їх візуалізації;

– відбір алгоритмів (визначення моделей для використання) та планування тестування моделей аналізу страхових випадків;

– побудова та тестування моделей страхових випадків згідно з обраними методами;

– оцінка результатів та технічний аналіз якості моделей страхових випадків;

– порівняння моделей, ухвалення рішення щодо можливості використання отриманих результатів.

1 ОГЛЯД ПРЕДМЕТНОЇ ГАЛУЗІ

1.1 Аналіз ринку страхування України

Згідно з чинним законодавством України поняття «страхування» набуває наступне значення: це такий вид цивільно-правових відносин, який захищає майнові інтереси фізичних та юридичних осіб у разі, коли настає подія, що визначена договором зі страхування або законом як страховий випадок, за рахунок грошових фондів, які сформовані сплатою особами страхових внесків (премій, платежів) та доходів цих фондів, що отримані шляхом розміщення їх коштів [1].

Страховий випадок – це подія, що відбулась та була передбачена страховим договором чи законодавством, та з моменту настання якої страховик зобов'язаний виплатити страхове відшкодування страхувальнику, застрахованій або іншій третій особі [1].

За організаційною складовою у структурі страхового ринку можна виділити такі основні категорії учасників:

– уповноважений орган державного нагляду за страховою діяльністю: Національний банк України, до липня 2020 року – Державна комісія з регулювання ринків фінансових послуг;

– страховики (страхові компанії, перестрахувальники, товариства взаємного страхування), їх об'єднання (Ліга страхових організацій України, страхові бюро, асоціації, спілки, пули), страхувальники (фізичні та юридичні особи), застраховані особи;

– посередники: страхові посередники (агенти, брокери), професійні оцінювачі збитків (аварійні комісари, аджастери, диспашери), професійні оцінювачі ризиків (андерайтери, сюрвеєри) [2].

Страхування представляє складну систему відносин, що структурується за формами, видами та галузями.

За формами, що мають правову основу, страхування в Україні буває добровільним та обов'язковим, які в свою чергу розділяються на низку окремих видів.

За сферами діяльності або спеціалізацією страховика виділяють страхування життя («life») та загальне (ризикове) страхування («non-life»).

Класифікація страхування за об'єктами виділяє три галузі:

а) особисте страхування (об'єкти – життя, здоров'я і працездатність страхувальників або застрахованих):

– страхування від нещасних випадків;

– медичне страхування;

– страхування життя і пенсій.

б) майнове страхування (об'єкти – майно в різних його видах: рухомі та нерухомі матеріальні цінності, грошові кошти, доходи):

– страхування майна громадян;

– страхування майна юридичних осіб.

в) страхування відповідальності (об'єкт – відповідальність за шкоду, завдану страхувальником життю, здоров'ю, майну третьої особи).

За організаційно-правовою формою страховика виділяють:

– комерційне страхування;

– взаємне страхування;

– державне страхування [3].

Деталізуючи об'єкти страхування, можна визначити його окремі види. Зазвичай, страхові компанії спеціалізуються на 1–2 галузях.

Страхові компанії та послуги, що вони надають, у сукупності формують страховий ринок, де в якості товару виступає страхова послуга (певний вид страхування) [4].

Отже, страховий ринок – це частина фінансового ринку, особлива соціально-економічна структура, сфера грошових відносин, де об'єктом купівлі-продажу є страховий захист, формуються пропозиція і попит на нього [3].

Первинна складова страхового ринку – це страхова компанія, де відбувається формування економічних відносини щодо укладення та обслуговування договорів страхування, а також утворюється та розподіляється страховий фонд [3].

Ринок страхування займає друге місце за рівнем капіталізації серед інших фінансових ринків у небанківському секторі. На рисунку 1.1 продемонстрована структура активів фінансового сектору України на перше півріччя 2021 року (авторська розробка за даними [5], [6]).

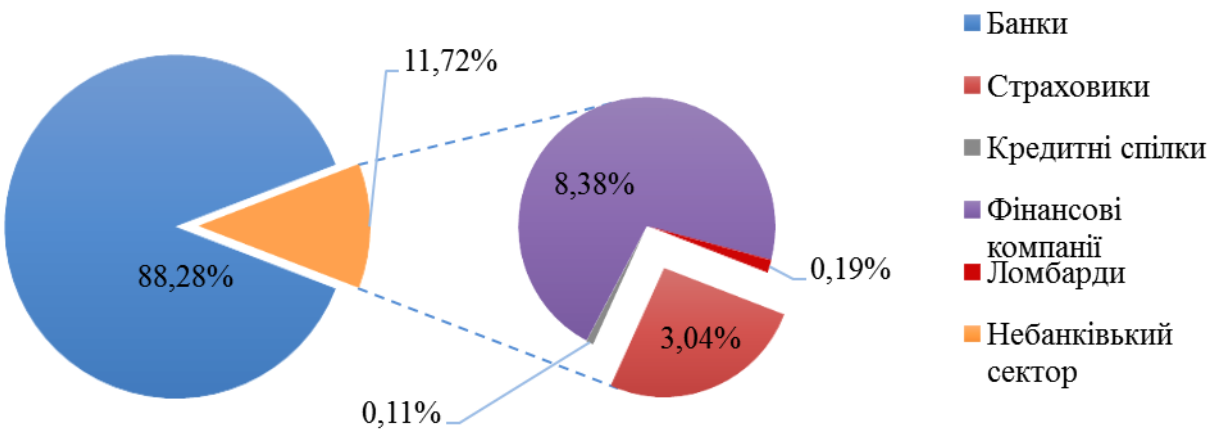

Рисунок 1.1 – Структура активів фінансового сектору України

Часто економічний розвиток країни за міжнародним досвідом визначається рівнем розвитку страхового бізнесу, дохідність якого вища промислового та банківського секторів у багатьох країнах, а також в умовах розвиненої ринкової економіки страхова справа являє собою механізм залучення інвестицій. На сьогодні ринок страхування України знаходиться у такому стані, коли його функціонування у фінансовій системі не є ефективним. За показником зібраних премій за останні роки загальний обсяг страхових послуг становить приблизно 0,06 % обсягу світу, що у порівнянні із США менше у 400 разів [7].

Показник відношення страхових премій до ВВП у багатьох країнах Східної Європи не перевищує 5%, а в країнах Західної Європи з добре

розвиненою сферою страхування він становить від 6,5 до 10,6 % [8].

У «Стратегії розвитку фінансового сектору України до 2025 року» одним з показників розвитку цієї галузі було визначено зростання частки страхових платежів у ВВП країни [5]. Станом на 2020 рік цей показник зменшився до 1,17% з 1,33% попереднього року. На рисунку 1.2 показана динаміка частки страхових виплат у ВВП з 2013 по 2020 роки, де простежується постійна тенденція до зменшення.

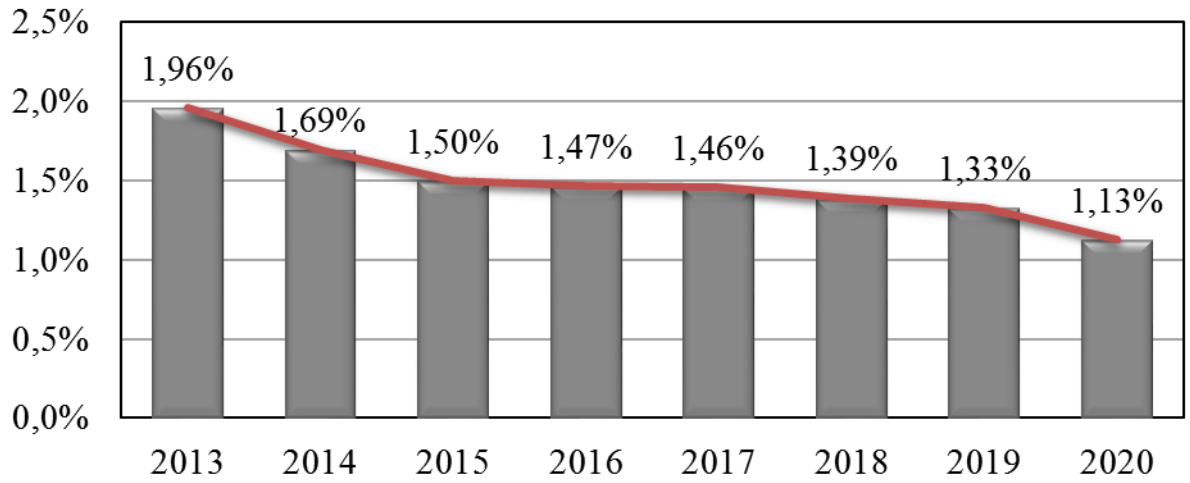

Рисунок 1.2 – Частка страхових платежів у ВВП за 2013–2020 рр.

Проаналізувавши показники діяльності ринку страхування України, можемо побачити, що обсяг активів страхового сектору у 2020 році зріс у порівнянні з попереднім роком, тем приросту складає 1,7 %.

Обсяг валових страхових премій знизилися на 14,7%, причиною цього послугувало припинення діяльності 23 компаній страховиків та зниження страхування ризиків. При цьому премії non-life-страховиків скоротились на 16,3%, а страховиків життя виросли на 8,4%. Співвідношення виплат до премій протягом 2020 року для страхування ризиків зберігалось на одному рівні та становило 35%, для життя – 13%.

Чисті страхові премії, що були отримані від страхувальників – фізичних осіб, становлять 57% структури надходжень, упродовж року

спостерігається їх ріст, що дорівнює 3%. Цей показник вказує, що населення зберігає попит на страхові послуги. У той же час зростають страхові премії від юридичних осіб. Загальне добровільне страхування залишається на стійкому рівні, його частка становить 80%. Основні показники діяльності страхового ринку України за період з 2016 по І півріччя 2021 року наведені у таблиці 1.1 (авторська розробка за даними [5], [6]).

Таблиця 1.1 – Показники діяльності страхового ринку України, млн.грн

	31.12.2016	31.12.2017	31.12.2018	31.12.2019	31.12.2020	І півріччя 2021	Темп приросту 2020/2019, %
1	2	3	4	5	6	7	8
Кількість зареєстрованих страховиків, у тому числі:	**310**	**294**	**281**	**233**	**210**	**181**	-9,87
- компанії зі страхування життя	39	33	30	23	20	19	-13,04
- інші	271	261	251	210	190	162	-9,52
Виключено з Державного реєстру за період	51	16	13	48	23	29	
Кількість укладених договорів страхування, за період, тис. од.	**179 471,20**	**185 482,90**	**201 077,50**	**196 923,70**	**120 576,50**	**63 169,52**	-38,77
Активи	**56 076,00**	**57 381,00**	**63 493,00**	**63 866,00**	**64 920,00**	**65 186,00**	1,65
Валові страхові премії	35 170,30	43 431,80	49 367,50	53 001,20	45 184,95	24 779,77	-14,75
Валові страхові виплати	8 839,50	10 536,80	12 863,40	14 338,30	14 852,70	8 703,26	3,59
Рівень валових виплат, %	**25,10**	**24,30**	**26,10**	**27,10**	**32,87**	**35,12**	
Чисті страхові премії (валові страхові премії за мінусом частки страхових премій, які сплачуються перестраховикам-резидентам)	26 463,90	28 494,40	34 424,30	39 586,00	40 350,20	23 479,74	1,93

Продовження таблиці 1.1

1	2	3	4	5	6	7	8
Чисті страхові виплати (валові страхові виплати за мінусом частки страхових виплат, які компенсовані перестраховиками-резидентами)	8 561,00	10 256,80	12 432,60	14 040,50	14 451,87	8 552,15	2,93
Рівень чистих виплат, %	**32,30**	**36,00**	**36,10**	**35,50**	**35,82**	**36,42**	

Показники обсягу виплат та рівня збитковості зростають у той час, коли обсяг премій зменшується, тому результати діяльності страхового бізнесу отримані за 2020 рік є одними з найгірших за останні роки.

За рівнем проникнення страхових послуг вищий показник спостерігається переважно у західних та південних регіонах України [5]. Кількість структурних підрозділів страховиків (джерело: НБУ [6]) наведена на рисунку 1.3.

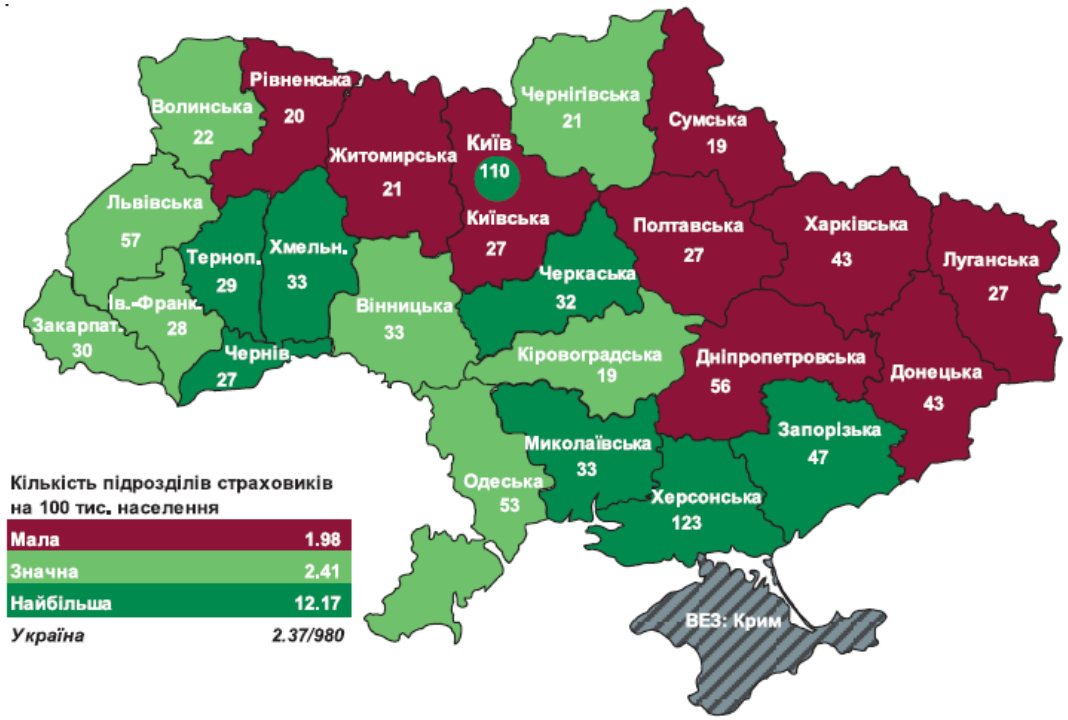

Рисунок 1.3 – Кількість структурних підрозділів страховиків у регіонах на перше півріччя 2021 року

Вагомою подією на українському ринку страхування у 2020 році та першому півріччі 2021 р. стало припинення діяльності значної кількості страхових компаній, а саме: за 2020 рік з державного реєстру було виключено 23 компанії, а за I півріччя 2021 – вже 29 одиниць. Згідно зі звітом НБУ станом на другий квартал 2021на ринку діють 181 зареєстрований страховик, з яких 19 – це компанії зі страхування життя, 162 – ризикового страхування.

Динаміка кількості укладених договорів має тенденцію до спаду, значення цього показнику у 2020 році знизилось на 38,7 % у порівнянні з попереднім роком. Динаміки кількості компаній та договорів (авторська розробка за даними [5], [6]) приведені на рисунку 1.4.

Рисунок 1.4 – Динаміка кількості страхових компаній та укладених договорів

На рисунку 1.5 (авторська розробка за даними [5], [6]) наведена динаміка кількості страхових компаній та їх активів за видами: «Life» та «Non-Life».

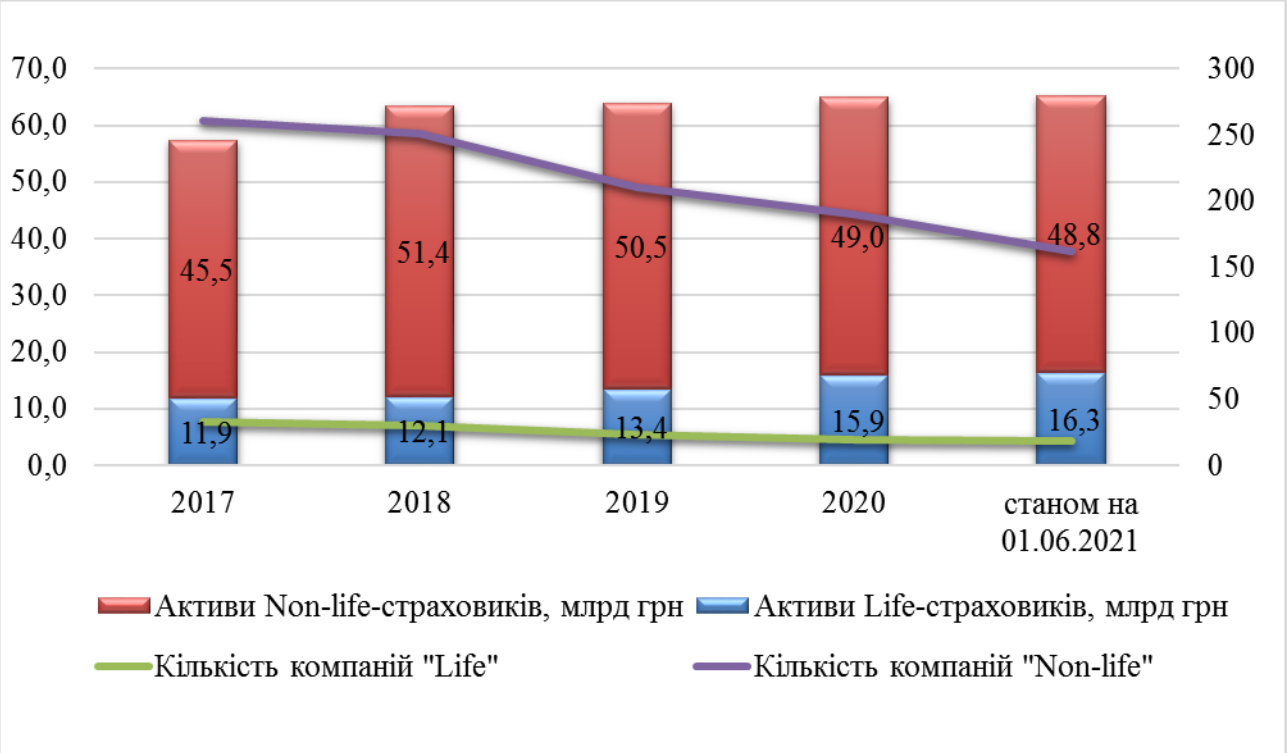

Рисунок 1.5 – Активи страхових компаній та їх кількість за видами

З продемонстрованого графіку помітно, що за двома видами страхування кількість компаній має тенденцію до зменшення. З 2020 року по І півріччя 2021 кількість Non-Life – страховиків скоротилась зі 190 до 162 одиниць, а Life – з 20 до 19. У той же час активи ризикового страхування скоротилися на 0,3%, а страхування життя збільшились на 2,75%.

Найпоширеніші види на українському ринку страхування:

– автострахування: КАСКО, ОСЦПВ, Зелена картка;

– особисте страхування: життя та медичне.

Посилаючись на думку експертів та їх розрахунки, частка платежів особистого страхування в Україні складає приблизно 4–5%, тоді як у Західній Європі та США цей показник становить 60%, а середнє значення у світі сягає 58,3% [7].

Структура премій та виплат за видами страхування у 2019, 2020 роках представлена на рисунку 1.6 (за даними НБУ, [6]).

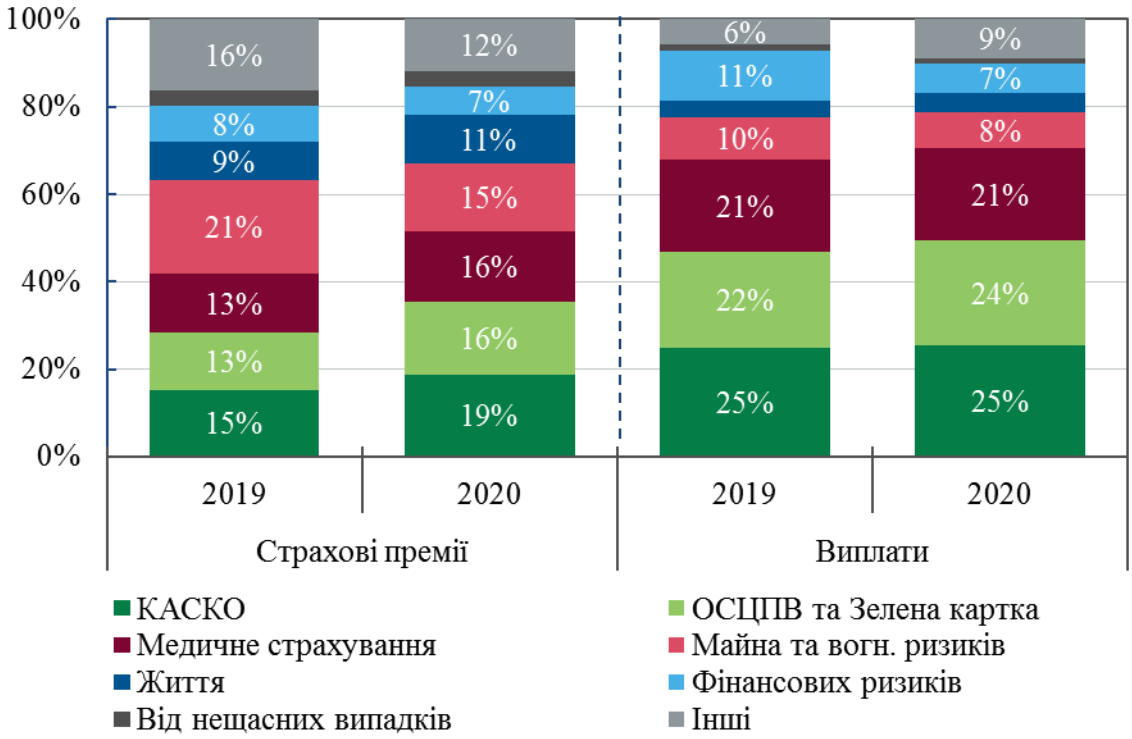

Рисунок 1.6 – Структура страхових премій та виплат за видами страхування у 2019, 2020 рр.

За 2020 рік обсяги премій за найпоширенішими видами зростали. Однак у страхуванні майна та вогневих ризиків премії зазнали значного зниження у порівнянні з попереднім 2019 роком. Ця зміна супроводжувалась припиненням діяльності лідерів даного виду та послугувала зміною структурі премій. Найбільша частка премій припадає на галузь автострахування.

На рисунку 1.7 представлено співвідношення премій та виплати за найпоширенішими видами страхування у II кварталі 2021 року.

На I півріччя 2021 лідерами за показником обсягів страхових премій стало автострахування та медичне страхування. Рівень обсягу страхових виплат за видами здебільшого залишився низьким.

Обсяги премій від добровільного виду страхування складають приблизно 75%, а рівень обсягів премій обов'язкового страхування збільшився на 30%, найзначніше збільшення спостерігається за ОСЦПВ [5].

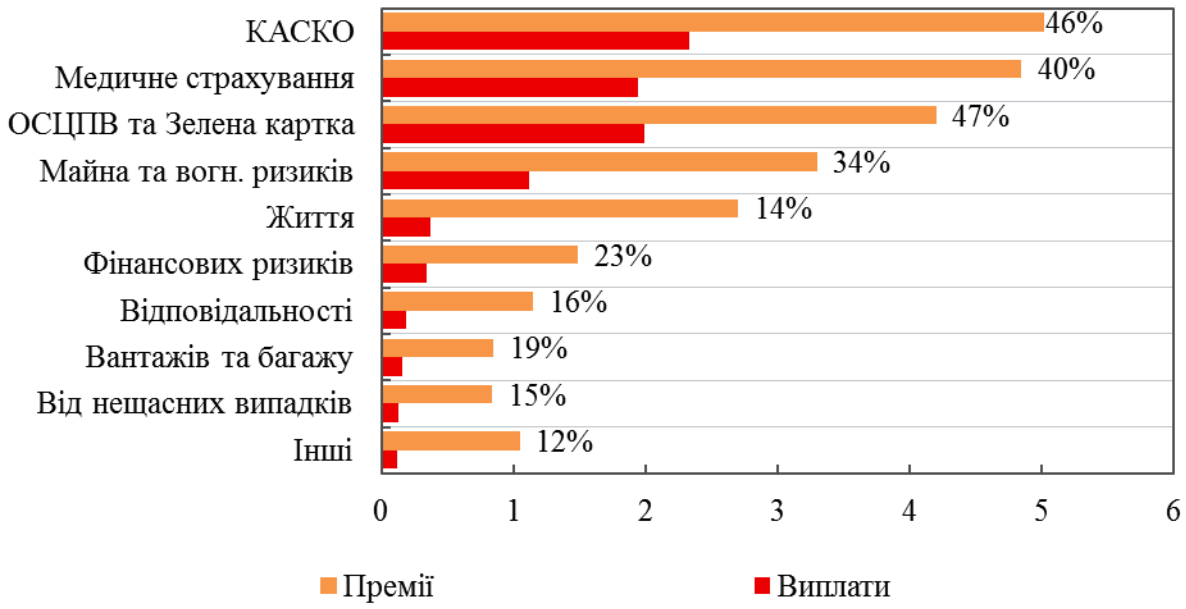

Рисунок 1.7 – Страхові премії та виплати за найпоширенішими видами страхування у II кварталі 2021 року, млрд грн

У II кварталі 2021 зросли коефіцієнти збитковості (loss ratio) для обох видів страхування: добровільного та обов'язкового. Рисунок 1.8 демонструє значення цього показника за окремими видами страхування.

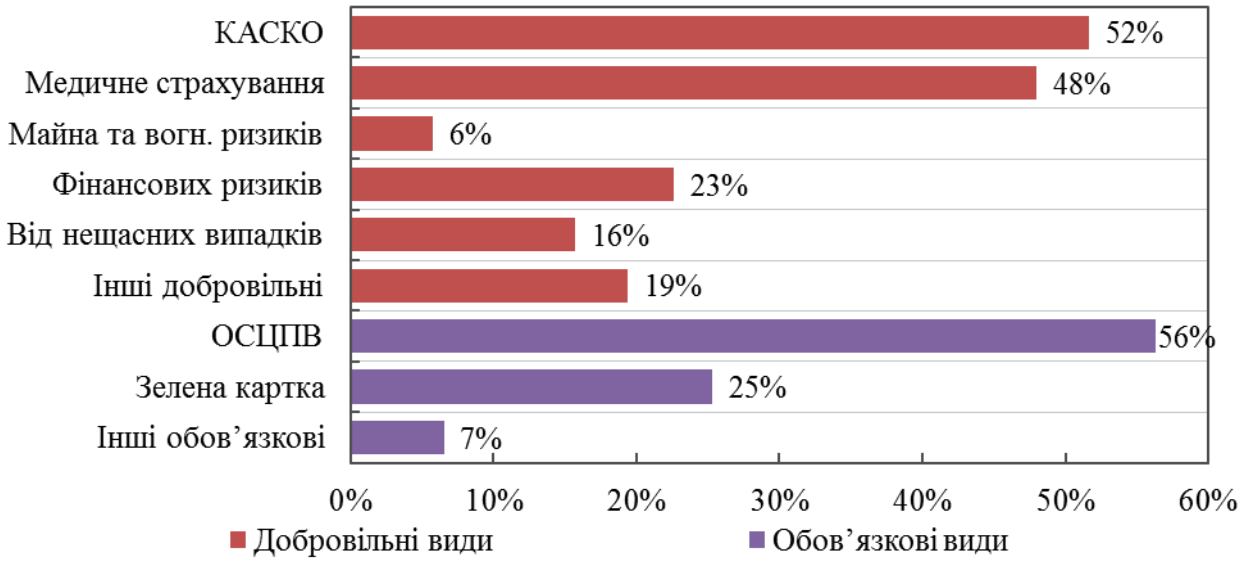

Рисунок 1.8 – Коефіцієнти збитковості (loss ratio) окремих видів страхування

Зазвичай цей показник сягає вищих значень для обов’язкових видів страхування, так найбільший коефіцієнт збитковості спостерігається за ОСЦПВ – 56%. Наступне місце серед найзбитковіших займають добровільні види: КАСКО – 52% та медичне страхування – 48%.

Обов’язкове особисте страхування від нещасних випадків на транспорті складає основну частку усіх договорів.

Отже, підводячи підсумок аналізу показників діяльності українського ринку страхування, рівень його розвитку можна оцінити як недостатній чи низький, він доволі сильно відстає від розвинених ринків європейських країн та США. Його частка у складі світового ринку дуже незначна, як низька і частка у ВВП країни. На думку експертів минулий рік показав одні з найгірших результатів діяльності сфери страхування за останні 10 років. Відмічається недовіра населення до страхування, причиною цього є ненадійність страхових компаній, які схильні до банкрутства через проведення неефективної інвестиційної політики. Для подальшого розвитку сфери страхування в Україні державі необхідно забезпечити ринок нормативною базою та своєю підтримкою [9].

1.2 Застосування методів машинного навчання у бізнес-цілях

Машинне навчання (Machine Learning, ML) – це підгалузь штучного інтелекту, що визначається самостійністю створення комп'ютером алгоритму дій, ґрунтуючись на певній моделі, яку задає людина, та поданих даних.

На сьогоднішній день машинне навчання знаходиться на піку свого розвитку, а його роль стрімко зростає. Так ML та штучний інтелект мають дуже вагомий вплив як на окремі найрізноманітніші сфери діяльності людини, так і на загальну світову економіку. Вже на цей час близько третини компаній в Америці, Європі та Китаї впровадили системи штучного інтелекту, що беруть за основу застосування методів машинного навчання.

Вважається, що за декілька років ці технології будуть слугувати майже у половині державних організацій та приватних компаній розвинених країн задля економії грошових ресурсів, покращення якості надання послуг, передбачення людської поведінки та вирішення низки інших можливих задач.

Машинне навчання користується великим попитом, доля якого постійно зростає, що пояснюється поширеністю галузей, у яких доцільне його застосування. Нові методи у цій галузі швидко розвиваються, і застосування машинного навчання розширилося майже до безмежних можливостей. Галузі, які залежать від величезних обсягів даних та потребують системи для їх ефективного та точного аналізу, визначили машинне навчання найкращим методом побудови моделей, розробки стратегії та планування.

Методи ML впроваджені у наступних сферах: реклама то просування для реалізації маркетингових задач, логістика, IT та IT-безпека, машинобудування та промислові галузі, сільське господарство, енергетика, туризм, охорона здоров'я та медицина, наукові дослідження, юриспруденція та фінансовий сектор, а також автоматизація процесів.

Практичне застосування машинного навчання сприяє досягненню бізнес-результатів, які можуть суттєво вплинути на чистий прибуток компанії. Для багатьох компаній, що вже використовують ці технології, вона не є інноваційною, а знаходиться на зрілій стадії, даючи можливість організаціям мати конкурентні переваги. Саме тому стрімко росте рівень інвестицій у цей напрямок, а кошти, що були вкладені, в значній мірі повертаються.

Вдаватися до методів машинного навчання у вирішенні питань бізнесу доцільно за наступних обставинах:

– наявний великий об’єм накопичених різноманітних даних, що потребують професійної обробки та використання машинних ресурсів;

– ці дані погано структуровані, мають пропущені або нетипові аномальні значення, не систематизовані;

– на перший погляд дані можуть бути позбавлені закономірностей та зрозумілих зв’язків, що зазвичай виявляються іншими аналітичними методами, натомість певний набір даних може бути корисним для визначення прихованих залежностей.

Методи машинного навчання спрямовані на допомогу у вирішенні багатьох бізнес-завдань.

Системи підтримки прийняття рішень допомагають керівництву передбачати тенденції, виявляти проблеми та прискорювати прийняття рішень, сприяючи подальшому розвитку організації. Тут мова йде про виконання задач прогнозування змінних та показників, що цікаві у конкретній ситуації, а також про виявлення відхилень, аномалій та залежностей. Це базові завдання, що можуть бути актуальними у багатьох бізнес-сферах. Наприклад, логістичні рішення включають алгоритми для пошуку факторів, які мають вплив на успіх ланцюгів поставок, управління активами та запасами, що дозволяє планувальникам оптимізувати процеси вибору, оцінки, маршрутизації та контролю якості постачань. У сфері охорони здоров'я інструменти машинного навчання допомагають лікарям визначати діагнози та варіанти лікування, підвищуючи ефективність догляду за хворими та покращуючи результати лікування пацієнтів, надається можливість прогнозування термінів життя пацієнтів із смертельними захворюваннями. У сільськогосподарській галузі технології ML на основі аналізу даних про кліматичні умови, ресурси та інші фактори допомагають фермерам приймати рішення з управління господарством [10].

Машинне навчання вирішує завдання автоматизації у популярній формі чат-ботів, що створюють комунікацію людини з машиною, виконуючи дії на основі запитів та вимог. Технології обробки природної мови (NLP) дозволяють вийти на інтерактивний рівень та реагувати на потреби користувачів, а розпізнавання аудіо-даних створює можливість

роботи віртуальних помічників на основі користувальницького вводу [11].

Рекомендаційні системи теж використовують ML для підвищення якості обслуговування клієнтів. Механізм аналізує дані по кожному клієнту, історії попередніх покупок та переглядів зі схожими вподобаннями для надання персоналізованих рекомендацій своїм потенційним користувачам продуктів та послуг. Ці технології широкого застосовуються у сфері маркетингу та продаж у розробці індивідуальних цільових маркетингових кампаній.

Одним із прикладів застосування методів класифікації та кластерізації є завдання з дослідження ринку, сегментації клієнтів за різними параметрами та аналіз їх складу. Машинне навчання використовується підприємцями для розуміння конкретних сегментів своєї клієнтської бази, щоб отримати уявлення про купівельні моделі певних груп покупців, що утворилась на схожих показниках (віку, доходів, статі, рівня освіти та інші). Це допомагає краще орієнтуватися на потреби кожного окремого сегменту та приймати рішення щодо утворення пропозицій. Методи кластерізації допомагають розподілити об’єкти за групами к випадках, коли взаємозв’язок явно не простежується [10].

Класифікація також застосовується у розпізнаванні зображень. Компанії вдаються до глибокого навчання та нейронних мереж, щоб зрозуміти зміст зображень. Ця технологія широко використовується: для відзначення фотографій у соціальних мережах, розпізнання облич, для виявлення службами безпеки злочинної поведінки, для візуального діагностування хвороб, виявленні ракових клітин та інше [10].

Фінансова сфера, до складу якої входить страховий бізнес, активно використовує методи ML. Вони допомагають відкалібрувати фінансові портфелі або оцінити ризики за кредитами та страховим андеррайтингом. Завдання штучного інтелекту у цій галузі включає можливість оцінювати хедж-фонди та аналізувати рух фондового ринку, щоб давати фінансові рекомендації.

Одним з основних варіантів використання машинного навчання у банківській сфері є боротьба із шахрайством. Машинне навчання найкраще підходить для цього варіанта використання, оскільки воно може сканувати величезні обсяги транзакційних даних та визначати наявність незвичайної поведінки. ML використовує набори даних, щоб визначати за секунди, які транзакції потрапляють у нормальний діапазон, а які транзакції виходять за рамки очікуваних норм, тобто є шахрайськими. Здатність машинного навчання розшифровувати закономірності і негайно виявляти аномалії, що виходять за рамки цих тенденцій, робить його чудовим інструментом виявлення шахрайських дій [11].

1.3 Вирішення завдань страхування методами машинного навчання

Технології штучного інтелекту, використання методів машинного навчання, зокрема глибокого навчання мають важливе значення для фінансової індустрії. Враховуючи високу конкуренцію на страховому ринку, передові рішення на основі даних пропонують конкурентні переваги за рахунок покращення управління ризиками, підвищення операційної ефективності, поліпшення обслуговування клієнтів, виявлення шахрайства на основі великої кількості даних, що були раніше зібрані компаніями.

Отже, досягнення машинного навчання слугують для вирішення наступних бізнес-завдань:

– покрашення страхового андеррайтингу. Встановлюючи ціни на контракти та послуги страховики покладаються на процес андеррайтингу, тобто на розрахунок ймовірності нещасного випадку та оцінки потенційного ризику для окремого клієнта. Конкретна ціна встановлюється, виходячи з аналізу даних про власника поліса, об'єкт покриття, статистики виникнення інцидентів в аналогічних ситуаціях та інших факторів;

– складніші алгоритми оцінювання. Оцінка ризиків є основою страхових компаній. Страховики можуть прийняти на себе більшість

ризиків, якщо вони знайдуть хорошу відповідність у цінах. Однак велика кількість компаній покладаються на традиційні методи при оцінці ризику. Індивідуальні клієнти часто оцінюються за допомогою застарілих показників, таких як кредитний рейтинг та історія збитків. Машинне навчання може запропонувати агентам нові інструменти та методи, що допомагають їм класифікувати ризики та розраховувати більш точні прогнозні моделі ціноутворення, що знижує коефіцієнти збитків [12];

– автоматизація та вдосконалення процесів обслуговування. Страховики опрацьовують тисячі претензій та відповідають на величезну кількість запитів клієнтів. Машинне навчання може покращити цей процес завдяки автоматизації переміщень заявки у системі від початкової реєстрації до аналізу та контакту з клієнтом. Крім делегування початкового спілкування з клієнтами чат-ботам, страховики можуть створювати більш персоналізовані та справедливі плани покриття збитків страхувальникам. У деяких випадках претензії можуть взагалі не вимагати роботи людей, що дозволяє їм більше часу приділяти більш складним претензіям. Таким чином клієнти зможуть отримати гарантовану оплату у разі втрати або пошкодження у найкоротший термін [12];

– точне прогнозування життєвої цінності клієнта. Життєва цінність клієнта (CLV) – показник, що визначає цінність клієнта для компанії як різницю між отриманим доходом та понесеними витратами. Страхові компанії прогнозують цей показник за допомогою даних щодо поведінки клієнтів, що дозволяє оцінити його потенційну прибутковість та створити більш персоналізовану маркетингову пропозицію. Моделі машинного навчання, засновані на поведінці, використовуються для прогнозування всіх критичних чинників майбутнього доходу компанії;

– забезпечення зростання продажів і більшого задоволення потреб клієнтів. ML поліпшує сегментацію клієнтів за соціальними та віковими факторами, інформацією про використовувані страхові продукти, взаємодією зі службою підтримки та іншим критеріями. Моделі машинного

навчання сприяють більш ефективним перехресним та додатковим продажам, дозволяючи компаніям збільшувати прибуток, виявляючи людей, які з великою ймовірністю куплять більше товарів одразу або згодом. За допомогою прогнозного моделювання страховики зможуть виявляти лояльних та нелояльних людей до певного типу полісів. Таким чином вони збільшують коефіцієнт конверсії продажів, пропонуючи правильні поліси конкретним покупцям, не витрачаючи час на тих, хто не зацікавлений у певних послугах [12];

– виявлення та запобігання шахрайству. Методи машинного навчання допомагають виявляти та розслідувати види шахрайства з боку агентів та клієнтів. Наприклад, деякі агенти зазвичай перебільшують факти про пошкодження автомобіля, щоб отримати вищий відсоток від компенсації. Клієнти ж, які хочуть швидко отримати гроші від страховика, можуть інсценувати автомобільну аварію чи крадіжку. ML допомагає захистити компанії та заощадити гроші при перевірці претензій. Інтелектуальні системи ідентифікують дані в різних страхових випадках, виявляють незвичайні або підозрілі зв'язки між різними клієнтами, пристроями та полісами. Вони також аналізують історичні дані про страхувальника та застраховане майно [13].

Запобігання шахрайству можливе ще під час реєстрації претензій. У цьому випадку компанії можуть виявити підозрілі стосунки між існуючими та новими клієнтами, номерами телефонів, IP-адресами, пристроями, банківськими рахунками, ремонтними майстернями та постачальниками медичних послуг [12].

Технології дозволяють страховикам отримувати дані про поведінку клієнтів та агентів та порівнювати їх із раніше зібраними даними про шахрайську поведінку. Методи машинного навчання допомагають виявляти агентів, які працюють не так, як інші, або помічати значні зміни у їхній звичній поведінці.

В процесі пошуку та вивчення робіт на тему виявлення шахрайства у

страхуванні було з’ясовано, що на вітчизняному просторі такі дослідження не проводяться чи не публікуються, тому їх популярність не розвинена. Однак є можливість ознайомитись з працями зарубіжних дослідників.

Hassan A.K.I. та Abraham A. у своїй роботі «Моделювання виявлення страхового шахрайства з використанням незбалансованої класифікації даних» пропонують інноваційний метод виявлення страхового шахрайства для усунення незбалансованого розподілу даних. Ідея заснована на побудовах моделей виявленого страхового шахрайства з використанням дерева рішень (DT), метода опорних векторів (SVM) та штучної нейронної мережі (ANN) на розділених даних, отриманих в результаті неповної вибірки (з заміною та без заміни) класу більшості та злиття с класом меншості. У статті використовується десятикратний метод перехресної перевірки (cross-validation). Кожен цикл крос-валідації полягає у випадковому розбитті відомого набору даних для використання однієї частини для навчання алгоритму, а іншої – для тестування та оцінки його ефективності. Цей процес повторюється кілька разів, а як показник ефективності навчання використовується середня помилка. Результати загальнодоступного набору даних щодо виявлення шахрайства при автострахуванні демонструють, що DT працює трохи краще, ніж інші алгоритми, тому модель DT використовувалася для порівняння між різними підходами до розділення з неповною вибіркою. Емпіричні результати показують, що запропонована модель дала найкращі результати [14].

Стаття Itri B., Mohamed Y., Mohammed Q., Omar B. присвячена оцінці ефективності та перевіреності найвідоміших алгоритмів ML для передбачення шахрайства. Автори застосували контрольований метод, який застосовувався до претензій щодо автомобільних даних анонімної страхової компанії. Вони запропонуваки підхід, який покращує релевантність результатів штучного інтелекту. Дослідження продемонструвало, що Random Forest працює краще серед усіх порівнюваних алгоритмів [15].

У статті Patel D. K. та Subudhi S. «Застосування Extreme Learning

Machine для виявлення шахрайства в автострахуванні» йдеться про розробку нової методології для виявлення аномальних претензій у записах автострахування за допомогою нейронної мережі Extreme Learning Machine (ELM). Автори пропонують наступний алгоритм дій: спочатку необроблений набір даних проходить процедуру попередньої обробки та розділяється на набори для навчання, перевірки та тестування. Потім генерується пул навчених класифікаторів ELM за допомогою різних комбінацій параметрів ELM у навчальному наборі. Обирається найкраща модель ELM, піддавши набір перевірці на навчених моделях. Після цього тестовий набір подається до перевіреної моделі для визначення характеру страхових претензій – законні чи зловмисні. Продуктивність моделі продемонстровано обширними тестами, проведеними на широко використовуваному наборі даних автострахування [16].

Pattanaik A. та Panigrahi S провели систематичне дослідження виявлення шахрайства при автострахуванні. Були визначені шахраї, їх основні види та підтипи відомих страхових шахрайств. Дослідники класифікували, порівняли та узагальнили майже всі опубліковані технічні та оглядові статті у цій галузі за останні 10 років. Вони запропоновали нову схему, в якій використовується метод вибору змінних на основі оптимізації рою часток (PSO) для вилучення нерелевантних та надлишкових ознак у наборі даних автострахування. Оскільки набір даних сильно спотворений за своєю природою, вони використали метод опорних векторів (QS-SVM), засновану методі вибірки для балансування даних. Після цього вони використовували дерево рішень (DT) та логістичну регресію (LR) для класифікації збалансованих даних. Ефективність запропонованої методології оцінюється експериментально з використанням набору реальних даних про шахрайство під час автострахування, взятих із літератури [17].

У статті Mubarek A. M. та Adalı E. «Техніка багатошарової нейронної мережі персептрона для виявлення шахрайства» надається визначення

системи виявлення вторгнень та її типів. Автори фокусуються на реалізації набору відомих алгоритмів класифікації машинного навчання (дерева рішень, наївний Байєсовський класифікатор штучні нейронні мережі), які можуть зменшити наявні недоліки системи виявлення вторгнення. Експериментальні результати на наборі даних дозволяють зробити висновок, що метод ANN-MLP (багатошаровий перцептрон) дає в середньому кращу продуктивність за рахунок обчислення «матриці неточностей», яка, у свою чергу, допомагає обчислити показники продуктивності, такі як «рівень виявлення точності», «точність» і «чутливість» [18].

У своїй роботі Ghorbani A. та Farzai S. після дослідження різних способів шахрайських злочинів у страхуванні використовували техніку кластеризації K-Means, щоб знайти моделі шахрайства в автомобільному страхуванні. Експериментальні результати вказують на високу точність у порівнянні зі статистичною інформацією, отриманою з наборів даних. Результати показують значні співвідношення між ефективними факторами у подібних випадках шахрайства [19].

Hanafy M. та Ming R., використовуючи дані реального життя, оцінили 13 підходів машинного навчання. Через незбалансовані набори даних прогнозування страхового шахрайства стало серйозною проблемою. Через те, що дані складаються здебільшого з класу «претензії без шахрайства» з невеликим відсотком «претензій з шахрайством». Таким чином, прогнозування шахрайства видається слабким з моделями класифікації; тому дане дослідження має на меті запропонувати підхід, який покращує результати алгоритмів машинного навчання, використовуючи методи повторної вибірки, такі як Random Over Sampler, Random under Sampler та гібридні методи, для вирішення проблеми незбалансованих даних. Після використання методів повторної вибірки ефективність усіх класифікаторів ML підвищується. Крім того, результати підтверджують, що не існує жодного методу, який би в цілому перевершував. Крім того, серед усіх

інших моделей класифікатор Stochastic Gradient Boosting отримав найкращий результат при використанні техніки гібридної передискретизації (over-sampling) [20].

В роботі Chun Yan та Yaqi Li досліджується технологія аналізу даних для боротьби з шахрайством із страхуванням автомобілів, було застосовано вдосконалений метод виявлення викидів, заснований на найближчому сусіді (KNN) з правилами відсікання, та вдосконалена модель ідентифікації шахрайства та відповідний алгоритм, а також правила асоціації використані для розробки закону шахрайства. Метод був перевірений під час експерименту, результати якого показують, що вдосконалений алгоритм ідентифікації шахрайства зі страхуванням автомобілів мав переваги: низьку часову складність, високий рівень розпізнавання, високу точність і низький вплив на K значення алгоритму [21].

Автори Chun Yan та Yaqi Li продовжують свою роботу та разом з Wei Liu та Maozhen Li розглядають використання методу випадкового лісу (Random Forest) для моделювання шахрайства. З огляду на те, що кількість вибірок у фактичних даних про відшкодування автомобільного страхування не є збалансованою, а кількість даних є великою, для створення моделі випадкового лісу вони відібрали реальні дані автомобільної страхової компанії. Дані були оброблені для перегляду індексу та проведено аналіз важливості кожної вхідної змінної до вихідної. Проаналізовано помилку моделі. Емпіричні результати показують, що: у порівнянні з традиційною моделлю, модель інтелектуального аналізу випадків шахрайства, яка використовує Random Forest, підходить для великих наборів даних і незбалансованих даних. Його можна краще використовувати для класифікації та прогнозування даних про відшкодування автомобільного страхування та визначення правил шахрайства. І він має кращу точність і надійність [22].

1.4 Висновки до розділу 1

У світовій практиці страхування було помічено, що приблизно в 10% усіх випадків спостерігається шахрайство або спроба обдурити страховиків. Однак для деяких страхових компаній цей відсоток не перевищує 5, а для якихось може досягати значення в 30%. Таку різницю можна пояснити тим, що перші можуть приділяти більше уваги і коштів для виявлення неблагонадійних клієнтів та впроваджувати сучасні рішення для визначення спроб шахрайства [13].

Через дії шахраїв страхові компанії зазнають дуже великих збитків. Так в розвинених країнах Європи та США річна сума збитків за усіма видами може становити близько 80 – 100 млрд дол. Звичайно ж український ринок страхування не так сильно потерпає від махінацій, це пояснюється низьким рівнем розвитку та невеликим попитом на страхові послуги. Тобто чим менше загальна кількість клієнтів, тим менша і частка серед них недобросовісних користувачів, бажаючих отримати страхові виплати обманним шляхом.

В межах української сфери страхування серед усіх випадків, за якими страхові компанії несуть збитки, майже 2% припадає на випадки з встановленим фактом шахрайства. Як у світі, так і в Україні основним видом страхування, де спостерігається найбільша кількість шахрайських випадків є автострахування, а саме: ОСЦПВ та КАСКО. В деяких європейських країнах на ці види припадає орієнтовно 80% усіх махінацій [13].

Через такі високі показники страховики вдаються до різноманітних методів аналізу, які здатні допомогти вирішити проблему захисту від шахраїв та таким чином запобігти великим грошовим втратам. Використання звичайних експертних методів затягує розгляд кожного випадку на тижні, в той час коли автоматизований процес може у 5 разів пришвидшити термін розслідування [5].

Тому страховим компаніям необхідно шукати сучасні рішення, які дозволять скоротити їх збитки шляхом виявлення спроб шахрайства. Таким рішенням можуть стати методи машинного навчання. Однак вітчизняні компанії, на відміну від іноземних, поки що мало забезпечені належними фахівцями чи ще не встигли запровадити технології машинного навчання. Не зважаючи на це страховики розуміють переваги інструментів ML та ефективність їх використання у цілях мінімізації збитків. Тому близько 90% організацій вважать корисним перехід до сучасних рішень у веденні бізнесу, а 15% навіть вже готові до використання методів машинного навчання, що дозволить отримати певні конкурентні переваги [5].

Великі компанії, які є лідерами на ринку страхових послуг, вже будують свою роботу по боротьбі з шахрайськими випадками, спираючись на результати моделювання, які з певної точністю класифікують клієнтські претензії на шахрайські чи ті, які такими не є; приймаючи до уваги отримані аналітичні звіти щодо виявлення існування зв'язків між різними факторами, які можуть бути непомітними при виконанні аналізу звичними способами. А інструменти візуалізації допомагають зробити наглядними приховані залежності, які важко виявити, дивлячись на цифри.

Отже, проблема виявлення шахрайства стає все більш поширеною з розвитком страхового ринку та несе серйозну загрозу для цієї галузі та фінансового сектору в цілому. Шахрайські схеми стають все більш складними та удосконаленими, тому потребують застосування аналізу, заснованого на методах машинного навчання, що допоможе за короткий термін обробити велику кількість даних, передбачити спробу шахрайства та таким чином запобігти небажаній втраті коштів.

2 АНАЛІЗ БАЗИ СТРАХОВИХ ВИПАДКІВ ТА ПОПЕРЕДНЯ ОБРОБКА ДАНИХ

2.1 Постановка та обґрунтування задач роботи

Для вирішення задачі аналізу страхових випадків методами машинного навчання була обрана база даних страхових претензій з автострахування, яка складається з інформації за 2 місяці 2015 року. Дані були взяті з онлайн-платформи Kaggle, що є середовищем з ML та спільнотою фахівців з Data Science.

У датасеті надана інформація щодо споживачів страхових послуг, деталі поліса (та умови договору страхування) та характеристика інциденту по кожній висунутій претензії. А головне – фіксація наявності чи відсутності шахрайства за позовами з відшкодування збитків.

Обрана база налічує 39 змінних та містить 1000 записів зі страхових претензій клієнтів. Опис змінних наведений у додатку А.

Виходячи з вмісту бази даних доцільною буде побудова моделей класифікації, завдання яких полягатиме в передбаченні шахрайства у страхових претензіях та в розподіли випадків на шахрайські чи ті, що такими не є. Проведений аналіз допоможе виявити приховані зв’язки та завдяки розумінню принципів, на яких формуються випадки, відстежити чинники, що визначають факт шахрайства.

Сформулюємо основні завдання, які вирішуються в ході роботи:

– первинний аналіз та передобробка даних, що включають опис бази, структури даних, їх розподілу, візуалізації ознак, очищення та підготовку до моделювання;

– побудова моделей класифікації на різних наборах змінних. Будуть використані такі методи: логістична регресія, метод опорних векторів, метод k-найближчих сусідів, байєсова класифікація, дерево рішень, випадковий ліс, класифікаційна нейронна мережа;

– оцінка та порівняння якості моделей класифікації, відбір найкращих класифікаторів.

Всі етапи поставлених задач реалізуються на високорівневій мові програмування Python, використовуючи інтерактивне середовище розробки Jupyter Notebook, яке дозволяє одразу бачити результат виконання коду, що стає можливим завдяки спеціальним бібліотекам та їх класам для аналізу даних, представленим на рисунку 2.1.

```
# Importing the libraries

# Для обробки даних
import numpy as np
import pandas as pd

# Для візуалізації
import matplotlib.pyplot as plt
import seaborn as sns

# Для моделювання та тестування
import sklearn
import imblearn
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.linear_model import LogisticRegression
from sklearn.metrics import confusion_matrix
import sklearn.metrics as metrics
from sklearn.metrics import roc_curve, auc
from sklearn.metrics import roc_auc_score
import statsmodels.api as sm
from sklearn.feature_selection import RFECV
from sklearn.svm import SVC
from sklearn.neighbors import KNeighborsClassifier
from sklearn.naive_bayes import GaussianNB
from sklearn.tree import DecisionTreeClassifier
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import mean_squared_error
from sklearn.preprocessing import MinMaxScaler
from sklearn.metrics import silhouette_score
from sklearn.cluster import KMeans
from sklearn.metrics.cluster import adjusted_rand_score
import scipy.cluster.hierarchy as sch
from sklearn.cluster import AgglomerativeClustering
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import StratifiedKFold

# Для роботи з штучними нейронними мережами
import keras
from keras.models import Sequential
from keras.layers import Dense
import SimpSOM as sps
```

Рисунок 2.1 – Імпорт необхідних бібліотек

2.2 Підготовка даних до моделювання

Першим кроком, з якого необхідно розпочинати дослідження страхових випадків, стане знайомство із базою, проведення первинного аналізу даних, з ціллю вивчення їх складу, структури, виявлення слабких та сильних сторін, розуміння особливостей, виходячи з яких робити висновки щодо побудови плану подальшої роботи.

Перш ніж завантажувати набір даних необхідним кроком буде імпортування бібліотек Python, завдяки яким здійснюється виконання коду.

Завантажуємо дані, що були надані у форматі CSV, та для наочності виводимо на екран результати, щоб переконатись, що все завантажилося належним чином. На рисунку 2.2 наведена частина бази, що вмістилась на екрані.

```
# Importing the dataset
df = pd.read_csv('data_di.csv', sep=',')
```

```
df
```

	months_as_customer	age	policy_number	policy_bind_date	policy_state	policy_csl	policy_deductable	policy_annual_premium	umbrella_limit	insured_zip
0	328	48	521585	2014-10-17	OH	250/500	1000	1406.91	0	466132
1	228	42	342868	2006-06-27	IN	250/500	2000	1197.22	5000000	468176
2	134	29	687698	2000-09-06	OH	100/300	2000	1413.14	5000000	430632
3	256	41	227811	1990-05-25	IL	250/500	2000	1415.74	6000000	608117
4	228	44	367455	2014-06-06	IL	500/1000	1000	1583.91	6000000	610706
...	...	...	...	...	...	...	...	...	...	...
995	3	38	941851	1991-07-16	OH	500/1000	1000	1310.80	0	431289
996	285	41	186934	2014-01-05	IL	100/300	1000	1436.79	0	608177
997	130	34	918516	2003-02-17	OH	250/500	500	1383.49	3000000	442797
998	458	62	533940	2011-11-18	IL	500/1000	2000	1356.92	5000000	441714
999	456	60	556080	1996-11-11	OH	250/500	1000	766.19	0	612260

1000 rows × 39 columns

Рисунок 2.2 – Частина первинно завантажених даних

У процесі ознайомлення з даними було виявлено, що змінні «collision_type», «property_damage» та «police_report_available» мають серед значень символ «?», що вказує на відсутність інформації за такими випадками по даним ознакам. Мова Python не розпізнає цей символ в якості

пропущеного значення, тому є необхідним замінити його на значення «NaN» та перерахувати кількість значень за змінними.

Виконаємо код для заміни, та за допомогою бібліотеки pandas виведемо інформацію про датафрейм, який представлений на рисунку 2.1.

Тепер можна побачити, що у перелічених раніше змінних не вистачає значень, а усі інші мають повний набір даних. Датафрейм складається з 18 змінних з числовим типом та 21 категоріальної змінної (object). Детальний перелік відношення ознак до певних типів надано на рисунку 2.3.

```
df = df.replace('?',np.NaN)
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 39 columns):
 #   Column                       Non-Null Count  Dtype
---  ------                       --------------  -----
 0   months_as_customer           1000 non-null   int64
 1   age                          1000 non-null   int64
 2   policy_number                1000 non-null   int64
 3   policy_bind_date             1000 non-null   object
 4   policy_state                 1000 non-null   object
 5   policy_csl                   1000 non-null   object
 6   policy_deductable            1000 non-null   int64
 7   policy_annual_premium        1000 non-null   float64
 8   umbrella_limit               1000 non-null   int64
 9   insured_zip                  1000 non-null   int64
 10  insured_sex                  1000 non-null   object
 11  insured_education_level      1000 non-null   object
 12  insured_occupation           1000 non-null   object
 13  insured_hobbies              1000 non-null   object
 14  insured_relationship         1000 non-null   object
 15  capital-gains                1000 non-null   int64
 16  capital-loss                 1000 non-null   int64
 17  incident_date                1000 non-null   object
 18  incident_type                1000 non-null   object
 19  collision_type               822 non-null    object
 20  incident_severity            1000 non-null   object
 21  authorities_contacted        1000 non-null   object
 22  incident_state               1000 non-null   object
 23  incident_city                1000 non-null   object
 24  incident_location            1000 non-null   object
 25  incident_hour_of_the_day     1000 non-null   int64
 26  number_of_vehicles_involved  1000 non-null   int64
 27  property_damage              640 non-null    object
 28  bodily_injuries              1000 non-null   int64
 29  witnesses                    1000 non-null   int64
 30  police_report_available      657 non-null    object
 31  total_claim_amount           1000 non-null   int64
 32  injury_claim                 1000 non-null   int64
 33  property_claim               1000 non-null   int64
 34  vehicle_claim                1000 non-null   int64
 35  auto_make                    1000 non-null   object
 36  auto_model                   1000 non-null   object
 37  auto_year                    1000 non-null   int64
 38  fraud_reported               1000 non-null   object
dtypes: float64(1), int64(17), object(21)
memory usage: 304.8+ KB
```

Рисунок 2.3 – Інформація по датафрейму

Здійснимо підрахунок кількості унікальних значень за факторами, результат якого наведено на рисунку 2.4. Можна побачити, що змінні «policy_number», «policy_bind_date», «insured_zip», «incident_location» мають у своєму складі більше 95% унікальних значень. Ці дані описують характеристику кожної окремої неповторної ознаки клієнта, договору чи інциденту. Змінна «incident_date» фіксує дату події, та як і інші не несе інформативності, але може сприяти погіршенню результатів аналізу та ускладненню процесу обробки і моделювання. Тому було прийняте рішення про виключення цих 5 ознак з датасета, після чого загальна кількість ознак складає 34.

```
df.apply(lambda x: len(x.unique()))
```

```
months_as_customer             391
age                             46
policy_number                 1000
policy_bind_date               951
policy_state                     3
policy_csl                       3
policy_deductable                3
policy_annual_premium          991
umbrella_limit                  11
insured_zip                    995
insured_sex                      2
insured_education_level          7
insured_occupation              14
insured_hobbies                 20
insured_relationship             6
capital-gains                  338
capital-loss                   354
incident_date                   60
incident_type                    4
collision_type                   4
incident_severity                4
authorities_contacted            5
incident_state                   7
incident_city                    7
incident_location             1000
incident_hour_of_the_day        24
number_of_vehicles_involved      4
property_damage                  3
bodily_injuries                  3
witnesses                        4
police_report_available          3
total_claim_amount             763
injury_claim                   638
property_claim                 626
vehicle_claim                  726
auto_make                       14
auto_model                      39
auto_year                       21
fraud_reported                   2
dtype: int64
```

Рисунок 2.4 – Виявлення кількості унікальних значень змінних

Наступним важливим кроком передобробки даних є виявлення пропущених значень та їх подальше усунення. За умови наявності пропусків у даних зменшується можливість виявлення їх реальних закономірностей, спотворюється статистична оцінка, що веде до появи помилок.

Виходячи з цього проведемо перевірку на наявність пропущених значень за усіма змінними, для цього запишемо і виконаємо наступний розділ коду, приведений на рисунку 2.5, результатом якого буде підрахована сума пропусків та їх відсоток у загальній кількості значень.

```
# Cheking Missing data
total = df.isnull().sum().sort_values(ascending=False)
percent = (df.isnull().sum()/df.isnull().count()).sort_values(ascending=False)
missing_data = pd.concat([total, percent], axis=1, keys=['Total', 'Percent'])
missing_data
```

	Total	Percent
property_damage	360	0.360
police_report_available	343	0.343
collision_type	178	0.178

Рисунок 2.5 – Перевірка кількості пропущених значень

Як і передбачалось раніше три змінні, а саме: «property_damage», «police_report_available» та «collision_type», мають пропуски 36%, 34,3% та 17,8% відповідно. Через невірну обробку пропущених значень зростає ймовірність виникнення помилок у моделях, що надалі призводить до неправильного прийняття рішень. А невдалий вибір методу заповнення пропусків в кращому випадку не покращить результати, а в гіршому може й значно понизити їх якість. Тому при виборі способу усунення пропусків розглянемо окремо кожну змінну з урахуванням її змісту.

Оскільки усі три ознаки мають категоріальний тип, то можна скористатися наступними способами позбавлення від пропущених даних:

– видалення записів, що містять пропуски;

– заміна модою, медіаною або іншим значенням [23].

Оскільки пропуски у наших даних складають від 17 до 36%, то використовуючи перший спосіб, ми позбавимося більше третини об'єму бази. Це призведе до того, що велика кількість інформації не буде використана, а подальші результати стануть менш репрезентативними. Тому у конкретній задачі будемо уникати реалізації такого методу. Спробуємо замінити дані іншими значеннями.

Для трьох факторів виконаємо підрахунок кількості значень за кожною категорією, щоб отримати уявлення, на які саме категорії розділяються значення та яким є розподіл даних за ними. Виведемо на екран результат, представлений далі на рисунку 2.6.

```
df['property_damage'].value_counts()

NO     338
YES    302
Name: property_damage, dtype: int64

df['police_report_available'].value_counts()

NO     343
YES    314
Name: police_report_available, dtype: int64

df['collision_type'].value_counts()

Rear Collision     292
Side Collision     276
Front Collision    254
Name: collision_type, dtype: int64
```

Рисунок 2.6 – Кількість значень по категоріям у змінних з пропусками

Змінні «property_damage» та «police_report_available» приймають значення «NO» або «YES», причому в обох випадках дані розподілені за категоріями майже порівну. Пропущення інформації у цих факторах означає відсутність пошкодження майна та поліцейського звіту, тому можна замінити пропуски на значення «NO» та приєднати таким чином до цієї категорії.

Змінна «collision_type» поділяється на три категорії типів зіткнення: «Rear Collision» (зіткнення ззаду), «Side Collision» (зіткнення з боку), «Front Collision» (зіткнення спереду). Дані за цими категоріями теж, як і у

попередніх випадках, розподілені майже у рівній кількості. Але ще 17,8% даних мають невизначений тип зіткнення. Можна було б скористатися методом заміни пропусків модою, однак найпопулярніша категорія «Rear Collision» налічує не набагато більшу кількість значень, ніж інші. Тому я вважаю недоцільним робити таку заміну, а натомість виділити окрему категорія типу зіткнення «Unknown» (невідомий) для типів, інформація за якими відсутня.

Виконаємо заміну пропущених значень та виведемо результат, продемонстрований на рисунку 2.7.

```
df['property_damage'].fillna('NO', inplace = True)
```

```
df['property_damage'].value_counts()
```

```
NO     698
YES    302
Name: property_damage, dtype: int64
```

```
df['police_report_available'].fillna('NO', inplace = True)
```

```
df['police_report_available'].value_counts()
```

```
NO     686
YES    314
Name: police_report_available, dtype: int64
```

```
df['collision_type'].fillna('Unknown', inplace = True)
```

```
df['collision_type'].value_counts()
```

```
Rear Collision     292
Side Collision     276
Front Collision    254
Unknown            178
Name: collision_type, dtype: int64
```

Рисунок 2.7 – Заміна пропущених значень та перевірка результатів

З приведеного рисунка можна побачити, що заміна відбулась належним чином. Повторна перевірка на наявність пропущених значень вказує, що такі випадки відсутні.

Для того, щоб побудувати збалансовану модель необхідно мати гладкий розподіл даних для навчання. Потрібно перевірити чи відповідають отримані дані нормальному закону розподілу [23].

Для безперервних даних наведемо статистичні показники: середнє значення, стандартне відхилення, мінімум, максимум, а також квартилі розподілу – кванти, кратні 25%, тобто відповідні 25% («нижній»), 50% («середній», тобто медіана) та 75% («верхній»).

Розраховані статистичні показники для неперервних даних приведені на рисунку 2.8.

	months_as_customer	age	policy_annual_premium	umbrella_limit	capital-gains	capital-loss	total_claim_amount	property_claim	injury_claim
count	1000.000000	1000.000000	1000.000000	1.000000e+03	1000.000000	1000.000000	1000.00000	1000.000000	1000.000000
mean	203.954000	38.948000	1256.406150	1.101000e+06	25126.100000	-26793.700000	52761.94000	7399.570000	7433.420000
std	115.113174	9.140287	244.167395	2.297407e+06	27872.187708	28104.096686	26401.53319	4824.726179	4880.951853
min	0.000000	19.000000	433.330000	-1.000000e+06	0.000000	-111100.000000	100.00000	0.000000	0.000000
25%	115.750000	32.000000	1089.607500	0.000000e+00	0.000000	-51500.000000	41812.50000	4445.000000	4295.000000
50%	199.500000	38.000000	1257.200000	0.000000e+00	0.000000	-23250.000000	58055.00000	6750.000000	6775.000000
75%	276.250000	44.000000	1415.695000	0.000000e+00	51025.000000	0.000000	70592.50000	10885.000000	11305.000000
max	479.000000	64.000000	2047.590000	1.000000e+07	100500.000000	0.000000	114920.00000	23670.000000	21450.000000

Рисунок 2.8 – Статистичні показники за неперервними даними

Близько 95% даних, що мають нормальний розподіл знаходяться всередині інтервалу, що обмежений двома стандартними відхиленнями в околиці середнього значення, а 99% попадають у межі трьох std. Значення, що надмірно віддалені від середнього, можуть мати серйозний вплив на якість моделювання [23].

Аналіз розрахованих показників вказує на наявність таких значень у змінних «policy_annual_premium», «umbrella_limit», «property_claim».

Перш ніж подавати дані на тренування моделей необхідно завершити їх передпідготовку та довести до прийнятного стану. Для цього здійснимо роботу з викидами за числовими змінними, для цього виконаємо функцію для пошуку та заміни даних, що виходять за межі 3 стандартних відхилень (рисунок 2.9).

Запишемо до датафрейму вже очищені від викидів змінні і повторно виведемо статистику, що представлена на рисунку 2.10.

```
# Function Outliers
def outliers(df):
    num_var = list(df._get_numeric_data().columns)
    for col_names in num_var:
        df[col_names] = df[col_names].apply(lambda y: df[col_names].mean()-3*df[col_names].std()
                        if y < df[col_names].mean()-3*df[col_names].std() else y)
        df[col_names] = df[col_names].apply(lambda y: df[col_names].mean()+3*df[col_names].std()
                        if y > df[col_names].mean()+3*df[col_names].std() else y)
    return(df)
```

```
# Outliers
df = outliers(df)
```

Рисунок 2.9 – Пошук та заміна викидів

	months_as_customer	age	policy_annual_premium	umbrella_limit	capital-gains	capital-loss	total_claim_amount	property_claim	injury_claim
count	1000.000000	1000.000000	1000.000000	1.000000e+03	1000.000000	1000.000000	1000.00000	1000.000000	1000.000000
mean	203.954000	38.948000	1256.476176	1.091898e+06	25126.100000	-26793.700000	52761.94000	7397.773749	7433.420000
std	115.113174	9.140287	243.570287	2.266996e+06	27872.187708	28104.096686	26401.53319	4818.993582	4880.951853
min	0.000000	19.000000	523.903965	-1.000000e+06	0.000000	-111100.000000	100.00000	0.000000	0.000000
25%	115.750000	32.000000	1089.607500	0.000000e+00	0.000000	-51500.000000	41812.50000	4445.000000	4295.000000
50%	199.500000	38.000000	1257.200000	0.000000e+00	0.000000	-23250.000000	58055.00000	6750.000000	6775.000000
75%	276.250000	44.000000	1415.695000	0.000000e+00	51025.000000	0.000000	70592.50000	10885.000000	11305.000000
max	479.000000	64.000000	1987.807651	7.993220e+06	100500.000000	0.000000	114920.00000	21873.748536	21450.000000

Рисунок 2.10 – Статистичні значення після позбавлення від викидів

В результаті виконання очищення, можемо спостерігати, що показники максимального, мінімального, середнього значень, а також стандартного відхилення за змінними «policy_annual_premium», «umbrella_limit», «property_claim» змінилися, що говорить про відсутність викидів.

Наш датасет складається з 18 категоріальних змінних. Більшість алгоритмів машинного навчання не можуть безпосередньо працювати із категоріальними змінними. Тому перед нами стоїть завдання перетворення текстових атрибутів на числові значення для можливості подальшої обробки.

Перед тим як переходити до кодування змінних необхідно скласти перелік усіх їх можливих значень. Запишемо код, який виводить кількість категорій за кожною змінною та їх значення. Виконання коду приведено на рисунку 2.11.

	col	nunique	unique	type
13	property_damage	2	[YES, NO]	object
0	policy_state	3	[OH, IN, IL]	object
1	policy_csl	3	[250/500, 100/300, 500/1000]	object
14	police_report_available	2	[YES, NO]	object
2	insured_sex	2	[MALE, FEMALE]	object
6	insured_relationship	6	[husband, other-relative, own-child, unmarried...	object
4	insured_occupation	14	[craft-repair, machine-op-inspct, sales, armed...	object
5	insured_hobbies	20	[sleeping, reading, board-games, bungie-jumpin...	object
3	insured_education_level	7	[MD, PhD, Associate, Masters, High School, Col...	object
7	incident_type	4	[Single Vehicle Collision, Vehicle Theft, Mult...	object
11	incident_state	7	[SC, VA, NY, OH, WV, NC, PA]	object
9	incident_severity	4	[Major Damage, Minor Damage, Total Loss, Trivi...	object
12	incident_city	7	[Columbus, Riverwood, Arlington, Springfield, ...	object
17	fraud_reported	2	[Y, N]	object
8	collision_type	4	[Side Collision, Unknown, Rear Collision, Fron...	object
16	auto_model	39	[92x, E400, RAM, Tahoe, RSX, 95, Pathfinder, A...	object
15	auto_make	14	[Saab, Mercedes, Dodge, Chevrolet, Accura, Nis...	object
10	authorities_contacted	5	[Police, None, Fire, Other, Ambulance]	object

Рисунок 2.11 – Перелік значень категоріальних змінних

Важливим моментом даного етапу є правильність вибору способу кодування значень. Спочатку спробуємо використати просте кодування, яке передбачає взяття для кожної категорії якогось числового значення. Експеримент показав, що на даних, закодованих таким чином, точність моделі класифікації нижча, ніж на даних, що закодовані іншим способом. Це може бути пояснено недоліком такого методу: модель може неправильно інтерпретувати закодовані категорії [23].

Пояснимо це на прикладі значень змінної «authorities_contacted», яка поділяється на 5 категорій, що закодовані від 0 до 4. Категорія «Police» приймає значення «0», а «Ambulance» – «4». Служба швидкої матиме більший вплив, ніж поліція, бо $4 > 0$; однак в реальності ці категорії не мають впорядкованості, тому ми матимемо спотворення.

Використаємо середнє кодування або цільове. Середнє кодування аналогічне кодуванню міток, за винятком того, що тут мітки безпосередньо пов'язані з ціллю [24]. Ідея полягає в тому, щоб замінити категоріальну змінну середнім значенням відповідної цільової змінної. Для кожної категорії обчислюється відповідне середнє значення мети (серед цієї категорії) та значення категорії замінюється цим середнім значенням. Іншими словами, щоб отримати цільове середнє значення, для кожної категорії змінної необхідно прорахувати кількість значень, що попадають в певний клас (ціль), та поділити на загальну кількість значень цієї категорії.

Реалізуємо цей метод за допомогою функції «groupby» (рисунок 2.12).

```
df[['authorities_contacted','fraud_reported']].groupby(['authorities_contacted'],
                as_index = False).mean().sort_values(by = 'fraud_reported', ascending = False)
```

	authorities_contacted	fraud_reported
3	Other	0.318182
0	Ambulance	0.290816
1	Fire	0.269058
4	Police	0.208904
2	None	0.065934

```
df['authorities_contacted'] = df['authorities_contacted'].replace(('None','Police','Fire','Ambulance','Other'),
                                                                   (0.066,0.209,0.269,0.291,0.318))
```

Рисунок 2.12 – Кодування категоріальних змінних

Таким чином закодуємо усі інші змінні, окрім результуючої. Змінна «fraud_reported» буде закодована методом кодування міток. Категорії «Y» (шахрайські випадки) присвоюємо значення «1», «N» – «0» (рисунок 2.13).

```
df['fraud_reported'] = df['fraud_reported'].replace(('Y','N'),(1,0))
df['fraud_reported'].value_counts()
```

```
0    753
1    247
Name: fraud_reported, dtype: int64
```

Рисунок 2.13 – Кодування результуючої змінної

Наступним кроком я вирішила замінити значення змінної «auto_year» на вік автомобіля. Для цього від 2015 року (дата інциденту) відняла рік авто, отримала нові значення та записала їх у «auto_year». Після цього змінна приймає значення від 0 до 20.

2.3 Візуалізація та оцінка статистичних характеристик бази

Для кращого розуміння складу даних та розподілу їх значень за класами проведемо вивчення кожної змінної та приведемо результати візуалізації для ознак, які є найбільш значущими у подальшому моделюванні.

Першою розглянемо ознаку «fraud_reported», яка виступає результуючою змінною та визначає факт наявності чи відсутності шахрайства у страхових претензіях. Вона представлена двома класами:

– «YES» у кількості 247 випадків (встановлений факт шахрайства);

– «NO» у кількості 753 випадків, де шахрайство не виявлено.

Слід відмітити, що ми маємо справу з незбалансованими даними: співвідношення класів дорівнює 1:4 (рисунок 2.14). Ця проблема та способи її вирішення будуть розглянуті далі.

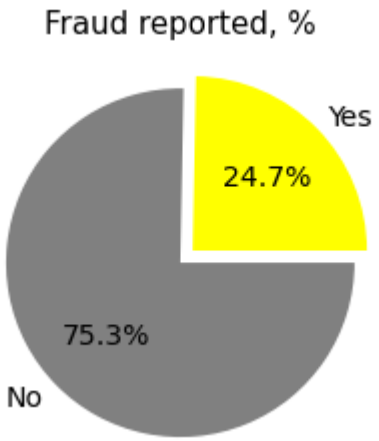

Рисунок 2.14 – Відсоткове співвідношення класів результуючої змінної «fraud_reported»

Перейдемо до розгляду характеристик деталей інциденту. На рисунку 2.15 у вигляді стовпчастих діаграм зображено розподіл значень змінної «incident_severity» (тяжкість інциденту).

fraud_reported	N	Y	Total
incident_severity			
Major Damage	109	167	276
Minor Damage	316	38	354
Total Loss	244	36	280
Trivial Damage	84	6	90
Total	753	247	1000

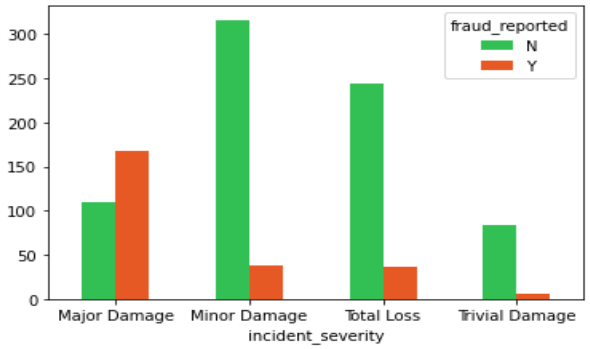

Рисунок 2.15 – Значення «incident_severity» за категоріями та класами

Ознака поділена на 4 категорії: велике пошкодження, менша шкода, повна втрата, незначна шкода. Чітко видно, що фіксація шахрайства у більшості випадків спостерігається за значенням «Major Damage», тобто коли тяжкість інциденту визначається як велике пошкодження, а для меншої шкоди («Minor Damage»), що є найчастішим значенням, характерна невелика доля шахрайства.

Тип інциденту теж приймає 4 значення: зіткнення кількох транспортних засобів, припаркована машина, одиночне зіткнення та викрадення авто (рисунок 2.16).

fraud_reported	N	Y	Total
incident_type			
Multi-vehicle Collision	305	114	419
Parked Car	76	8	84
Single Vehicle Collision	286	117	403
Vehicle Theft	86	8	94
Total	753	247	1000

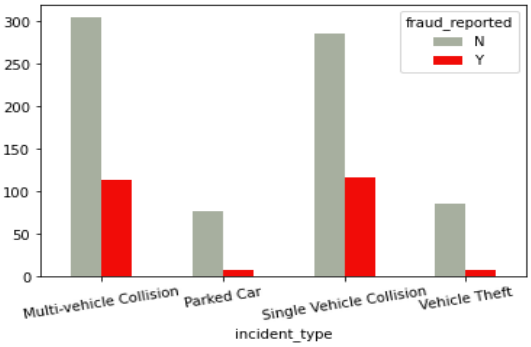

Рисунок 2.16 – Значення змінної «incident_type» за категоріями та класами

Для «incident_type» шахрайські дії частіше відмічаються при зіткнення кількох авто та при одиночному зіткненні, які є найпопулярнішими значеннями.

Контакт з органами влади притаманний обом класам (рисунок 2.17). Категорія «Police» нараховує більшу кількість значень.

fraud_reported authorities_contacted	N	Y	Total
Ambulance	139	57	196
Fire	163	60	223
None	85	6	91
Other	135	63	198
Police	231	61	292
Total	753	247	1000

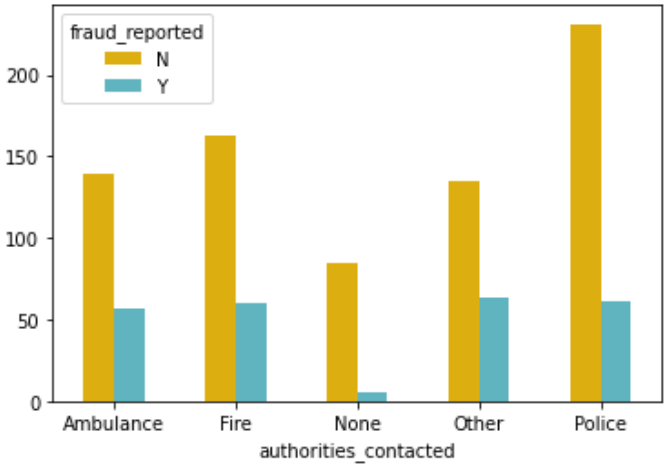

Рисунок 2.17 – Значення змінної «authorities_contacted

Для шахрайських випадків найпопулярнішим типом зіткнення є зіткнення ззаду («Rear Collision»), другий нешахрайський клас має майже рівномірний розподіл значень (рисунок 2.18).

fraud_reported collision_type	N	Y	Total
Front Collision	184	70	254
Rear Collision	201	91	292
Side Collision	206	70	276
Unknown	162	16	178
Total	753	247	1000

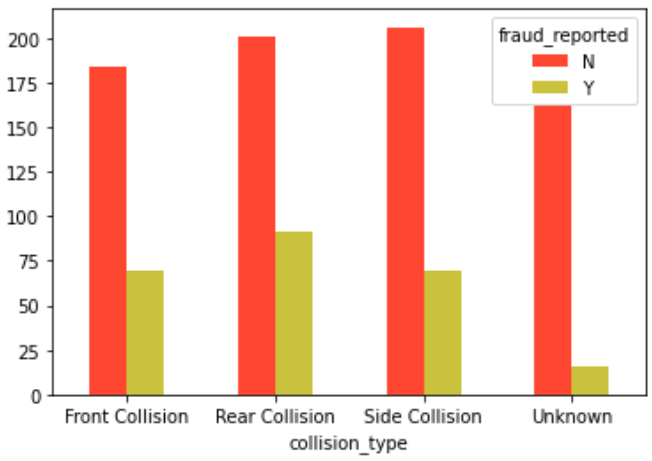

Рисунок 2.18 – Значення змінної «collision_type» за категоріями та класами

Претензії з зафіксованим фактом шахрайства переважно характеризуються відсутністю пошкодження майна та відсутністю поліцейського звіту (рисунок 2.19 – 2.20) демонструють розподіли значень за цими характеристиками.

fraud_reported	N	Y	Total
property_damage			
NO	529	169	698
YES	224	78	302
Total	753	247	1000

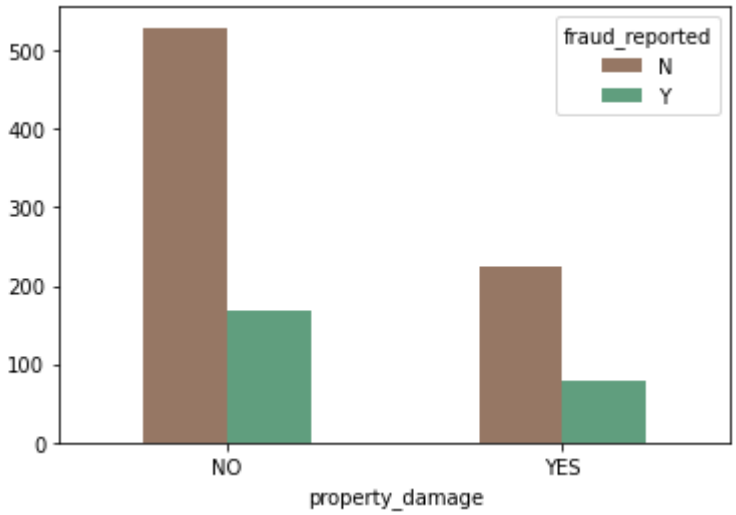

Рисунок 2.19 – Значення «property_damage» за категоріями та класами

Загальна кількість значення «No» по кожній змінній сягає приблизно 70% за двома класами.

fraud_reported	N	Y	Total
police_report_available			
NO	511	175	686
YES	242	72	314
Total	753	247	1000

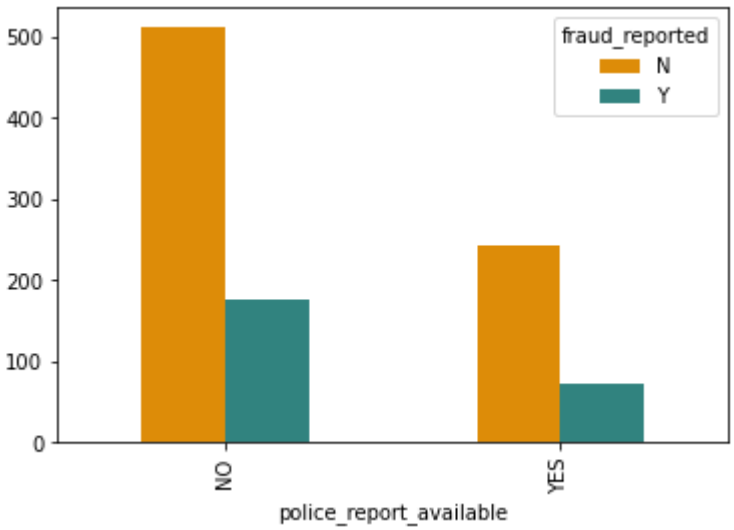

Рисунок 2.20 – Кількість значень змінної «police_report_available»

Ще однією серед значущих змінних, описуючих інцидент, є модель автомобіля, яка поділяється на 39 значень. У шахрайських претензіях найбільш популярні моделі: «RAM», далі – «A3», «F150», «Jetta». Для класу відсутності шахрайства «Wrangler» виступає найчастішою моделлю (рисунок 2.21).

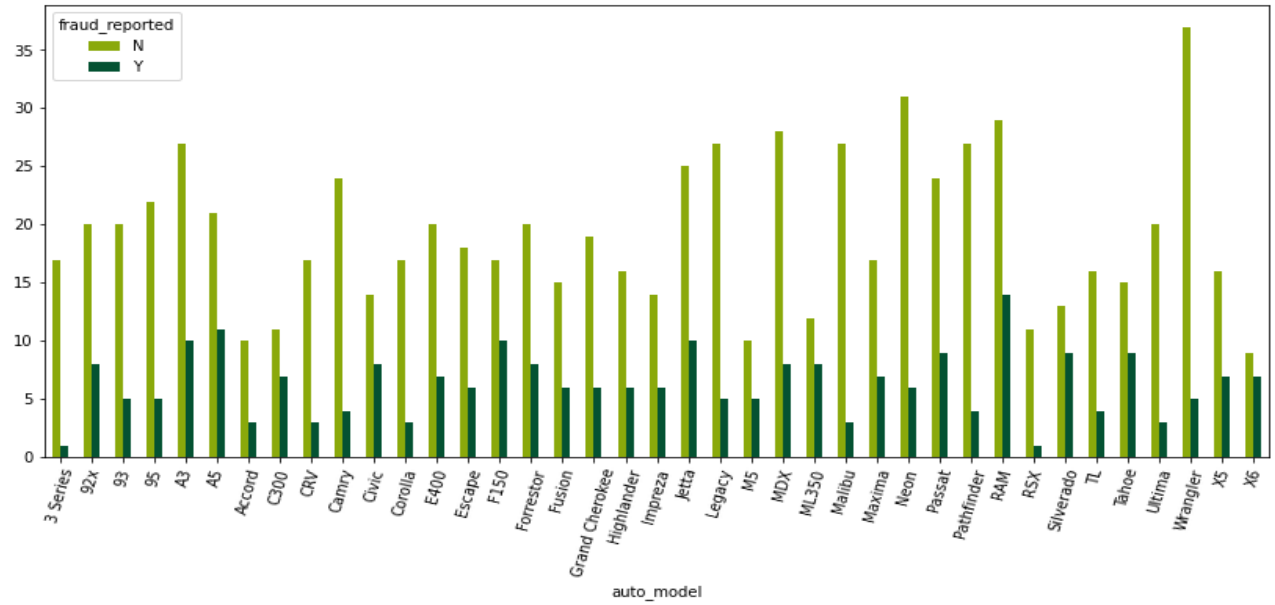

Рисунок 2.21 – Кількість значень змінної «auto_model»

Найчастіші значення змінної «auto_make» (усього 14 категорій) для класу шахрайства: Ford, Mercedes, для «чесного» класу – Nissan, Saab.

Перейдемо до розгляду якісних характеристик клієнтського складу. Найбільш чітко виражену приналежність до класів має змінна «insured_hobbies», яка розподілена на 20 категорій різних хобі клієнтів, значення за категоріями та класами продемонстровані на рисунку 2.22.

Можемо побачити, що для клієнтів-шахраїв найбільш характерні два види хобі: «chess», «cross-fit». А позитивні клієнти мають наступні хобі: «camping», «golf», «kayaking».

На рисунку 2.23 зображено розподіл за 14 видами зайнятості страхувальників.

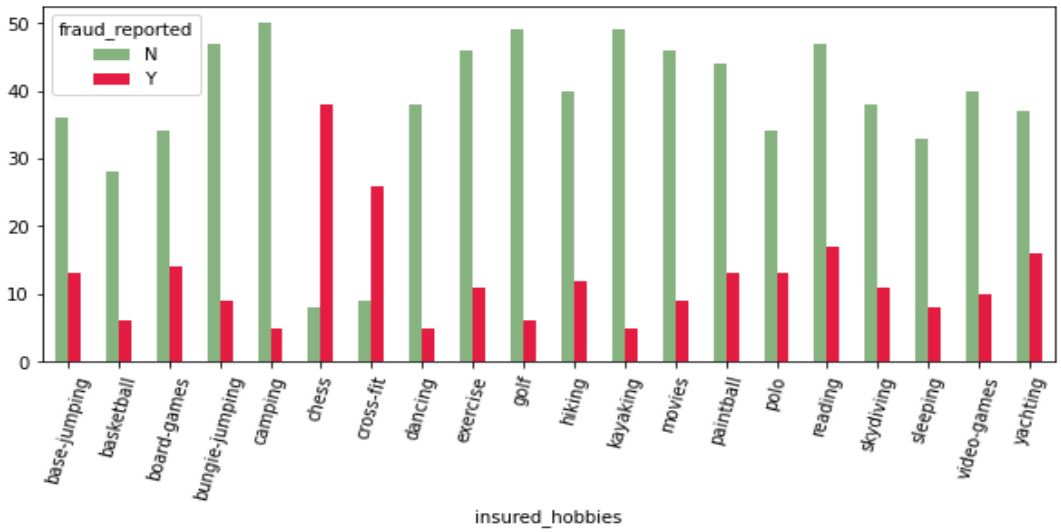

Рисунок 2.22 – Кількість значень змінної «insured_hobbies»

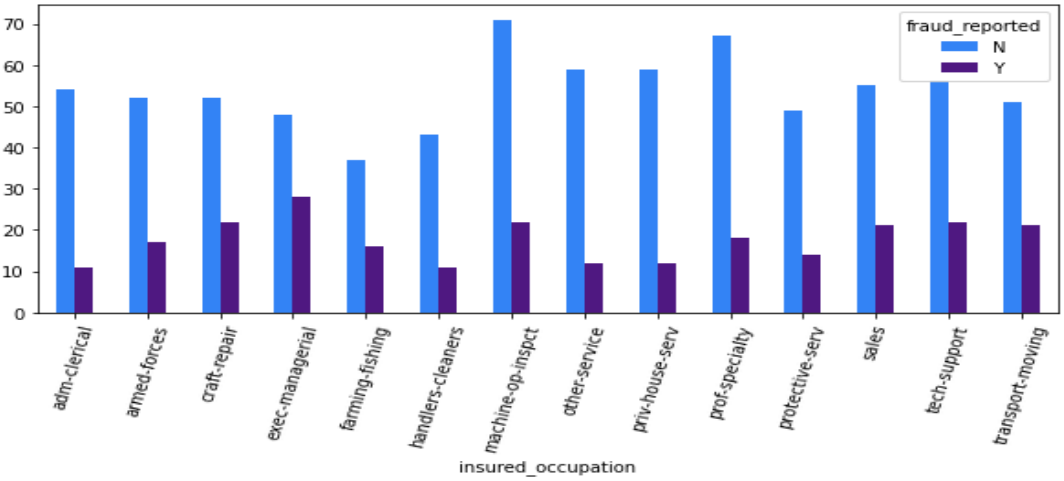

Рисунок 2.23 – Кількість значень змінної «insured_occupation»

Більша кількість шахрайських випадків спостерігається по клієнтам, що зайняті як «exec-managerial» (топ – менеджер). А для претензій з відсутністю шахрайства характерні такі види зайнятості: «machine-op-inspct» (машиніст-оператор), «prof-specialty» (профільний фахівець).

Змінна сімейного стану клієнтів по шахрайським позовам найчастіше приймає значення «other-relative» та «not-in-family», а для класу відсутності шахрайства найбільшу кількість значень має категорія «own-child» (має дитину). Рисунок 2.24 відображає розподіл змінної «insured_relationship».

fraud_reported	N	Y	Total
insured_relationship			
husband	135	35	170
not-in-family	129	45	174
other-relative	125	52	177
own-child	144	39	183
unmarried	107	34	141
wife	113	42	155
Total	753	247	1000

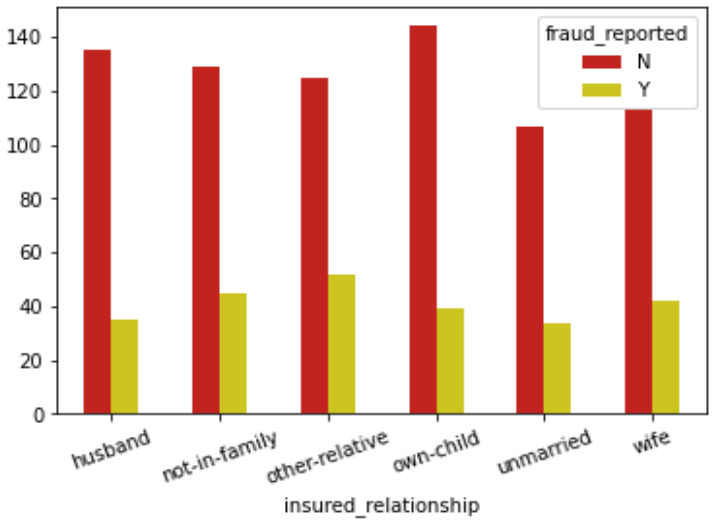

Рисунок 2.24 – Кількість значень змінної «insured_relationship»

Перейдемо до розгляду числових змінних, для чого побудуємо графіки їх розподілу у вигляді гістограм, які продемонстровані на рисунку 2.25.

Одразу можемо побачити, що наші 16 змінних поділяються на:

– безперервні : «months_as_customer», «age», «umbrella_limit», «capital-gains», «capital-loss», «total_claim_amount», «injury_claim», «property_claim», «vehicle_claim»;

– дискретні: «policy_deductable», «incident_hour_of_the_day», «number_of_vehicles_involved», «bodily_injuries», «witnesses», «auto_year».

Проаналізуємо дискретні змінні.

Змінна «policy_deductable» (франшиза поліса) це обумовлена частина коштів, яку страхова компанія не відшкодовує, тобто віднімає від суми виплати. Вона представлена трьома значеннями: «500», «1000», «2000», які розподілені за класами претензій майже рівномірно.

Ознака «incident_hour_of_the_day» приймає 24 значення, але залежності шахрайства від часу не розкриває, тому не є інформативною.

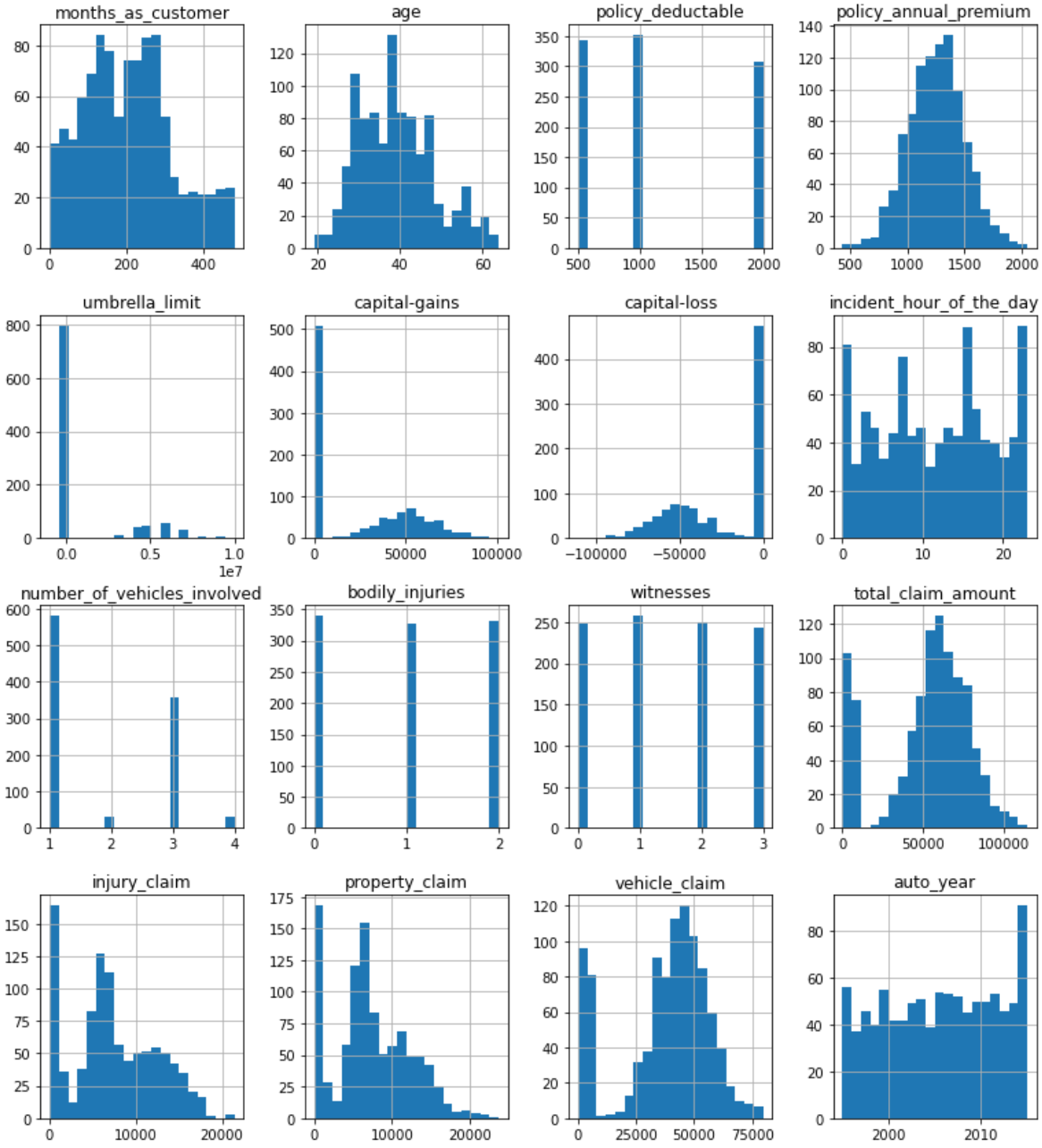

Рисунок 2.25 – Розподіл даних числових змінних

Кількість задіяних транспортних засобів має чотири значення: від 1 до 4 (рисунок 2.26).

Для шахрайських випадків найбільш характерне задіяння одного транспортного засобу та трьох.

Характеристика «bodily_injuries» (ступінь тілесних ушкоджень) приймає значення «1», «2», «3». Шахрайство незначно більше фіксується при значенні «3».

Змінна «witnesses» представлена чотирма значеннями від 0 до 3, шахрайські претензії найчастіше приймають значення «2», тобто інцидент має двох свідків.

Дані по «auto_year» приймають значення від 1995 до 2015 року. Авто 2004, 2007, 2011, 2013 років налічують найбільшу кількість шахрайських випадків.

fraud_reported	N	Y	Total
number_of_vehicles_involved			
1	448	133	581
2	21	9	30
3	264	94	358
4	20	11	31
Total	753	247	1000

Рисунок 2.26 – Кількість значень змінної «number_of_vehicles_involved»

2.4 Висновки до розділу 2

Обрана база налічує 1000 випадків та 39 показників за страховими випадками. Змінні представлені двома типами: 18 є числовими, а інші 21 – категоріальними. В процесі обробки з датасету було видалено 5 неінформативних ознак з великою кількістю унікальних значень. У трьох змінних («property_damage», «police_report_available», «collision_type») знайдені пропущені значення, та виконано їх заміну. Перевірка на наявність викидів показала, що вони містяться у змінних «policy_annual_premium», «umbrella_limit» та «property_claim», після чого для них була проведена заміна. Категоріальні змінні були закодовані з використанням середнього цільового кодування.

Побудована візуалізація надає можливість побачити залежність між

значеннями показників та їх приналежністю до класів. Це дозволяє скласти уявлення про те, на підставі яких значень факторів формується страховий випадок класу шахрайства або відмінний від нього.

Тепер коли усі дані майже підготовані, виведемо кореляційну матрицю, яка допоможе первинно оцінити значущість факторів (рисунок 2.27).

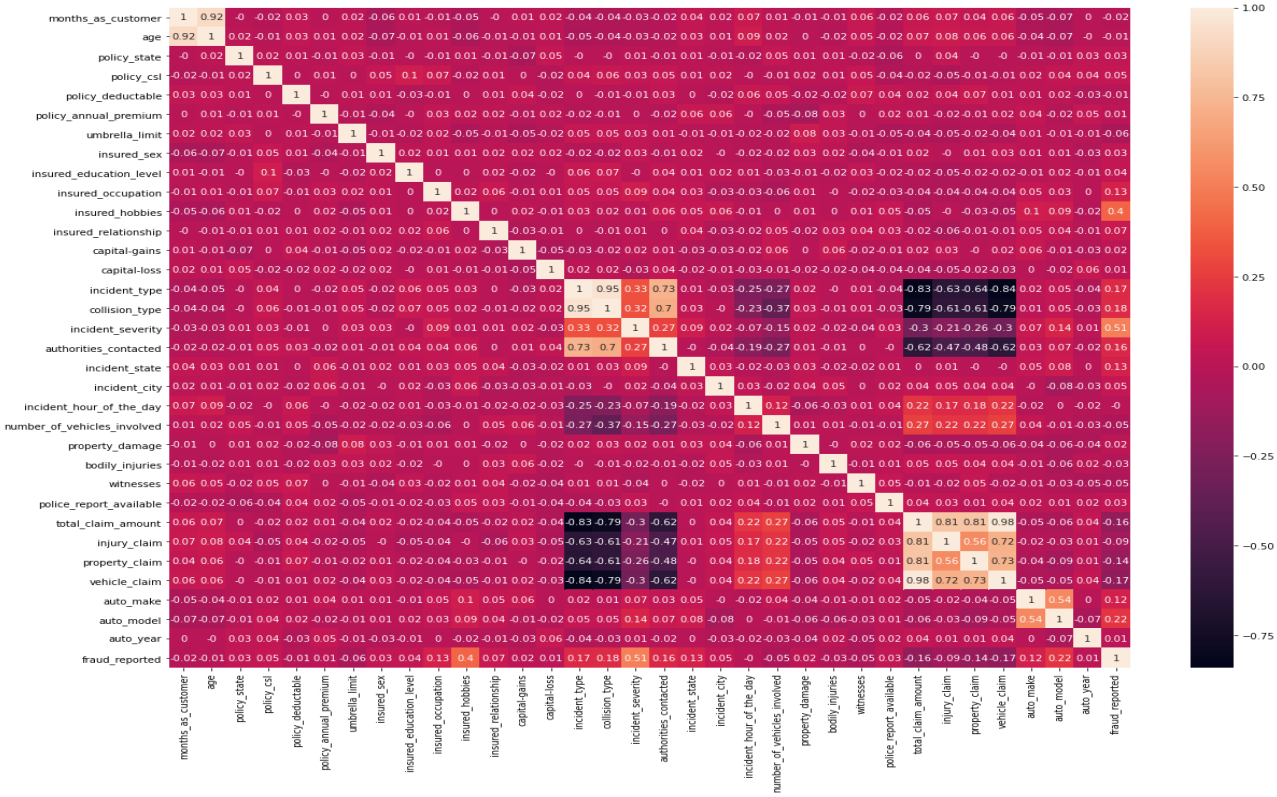

Рисунок 2.27 – Кореляційна матриця ознак

Найбільш значущими змінними є: «insured_occupation», «insured_hobbies», «incident_type», «collision_type», «incident_severity», «authorities_contacted», «auto_make», «auto_model», що підтверджує попередньо проведений опис.

Після проведених етапів обробки датасета, ми отримали сформований набір підготовлених даних, який можна далі використовувати для моделювання.

3 ПОБУДОВА ТА ТЕСТУВАННЯ МОДЕЛЕЙ КЛАСИФІКАЦІЇ СТРАХОВИХ ВИПАДКІВ

3.1 Вирішення проблеми незбалансованої вибірки

Будуючи модель класифікатора необхідно визначити, яка подія буде позитивною, а яка негативною. Так як наша задача полягає у виявленні шахрайських випадків, то відповідно позитивною подією буде наявність факту шахрайства (клас 1), а негативною – його відсутність (клас 0).

Перед переходом до процесу безпосереднього моделювання необхідно визначитись з розподілом даних на набори для навчання та тестування.

Для оцінки достовірності моделей будемо користатися одноразовою перехресною перевіркою, яка реалізується розбиттям вибірки на взаємодоповнювані підвибірки: навчальну (training) та тестову (testing). Перша слугує для тренування моделей, а друга для оцінки результатів [23].

Розділимо підготовані дані на підвибірки, виділивши на навчання 70% випадків, та здійснимо перевірку такого розподілу (рисунок 3.1).

```
from sklearn.model_selection import train_test_split
train, test = train_test_split(df, test_size=0.3, random_state=256)
```

```
pd.Series(train['fraud_reported']).value_counts()
```

```
0    525
1    175
Name: fraud_reported, dtype: int64
```

```
pd.Series(test['fraud_reported']).value_counts()
```

```
0    228
1     72
Name: fraud_reported, dtype: int64
```

Рисунок 3.1 – Розбиття даних на набір для навчання та тестування (70/30)

Як і було зауважено раніше, ми маємо справу з незбалансованою вибіркою, і тестовий, і навчальний набір мають розподіл класів 75% на 25%. Клас шахрайських випадків, який цікавить нас найбільше, виступає класом меншості.

Незбалансовані набори даних поширені в багатьох областях і секторах, і, звичайно ж, це включає фінансові послуги. Фахівці з обробки даних стикаються з ними у багатьох контекстах – від шахрайства до проблемних кредитів. Алгоритми машинного навчання намагаються ідентифікувати ці поодинокі випадки у досить великих наборах даних. Через невідповідність класів у змінних алгоритм має тенденцію відноситись до класу більшості, водночас даючи помилкове відчуття високоточної моделі. Нездатність передбачити рідкісні події, що належать до меншості, так і точність, що вводить в оману, відволікають від побудованих прогностичних моделей [26]. Навчання моделей на таких даних може призвести до неправильної класифікації та великої кількості помилкових визначень, бо модель буде орієнтуватися на виявленні негативних випадків, у той час, коли нас цікавлять саме позитивні.

Щоб уникнути цієї проблеми, необхідно на навчання моделі подати набір даних з рівномірним розподілом класів чи перебалансувати навчальну вибірку таким чином, щоб якомога більше значень класу шахрайства потрапила на тренування моделей. Скористаємося цими способами перебалансування та створимо ще два варіанти розбиття вибірки.

Існує відомий варіант роботи з незбалансованими даними, який називається ресемплінг (передискретизація). Його суть [27] полягає у видаленні елементів з занадто великого набору (андерсемплінг) та/або додаванні більшої кількості елементів в недостатньо великий набір (оверсемплінг). Принцип роботи цього способу представлено на рисунку 3.2.

У нашій ситуації, коли датасет складається з досить невеликої кількості випадків, користатися андерсемплінгом є недоцільним, тому використовувався метод оверсемплінгу.

Розподіл класів навчального набору даних до проведення балансування вибірки наведено на рисунку 3.3.

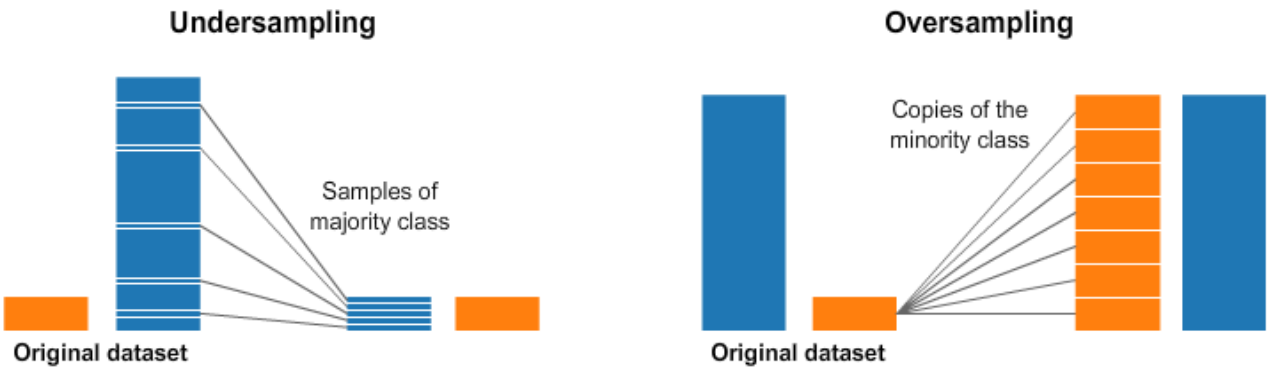

Рисунок 3.2 – Принцип роботи передискретизації

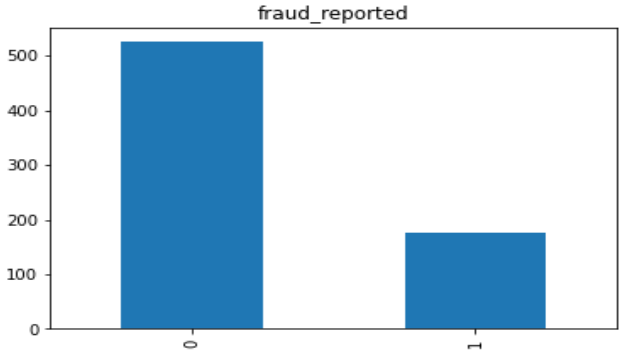

Рисунок 3.3 – Розподіл класів навчального набору даних

Найпростішим способом реалізації надмірної вибірки є дублювання випадкових записів із класу меншості. Слід зазначити, що недоліком такої вибірки може стати перенавчання моделі, що може призвести до поганого узагальнення тестового набору.

Застосування рандомного оверсемплінгу дозволяє збалансувати класи навчальної вибірки, збільшивши її на 450 шахрайських випадків (рисунок 3.4).

Іншим варіантом оверсемплінгу є використання методу SMOTE (Спосіб Передискретизації Синтезованих Меншин), який створює елементи в безпосередній близькості від тих, що вже існують у меншому наборі.

Для SMOTE обирається кілька спостережень та використовується міра відстані для синтетичного створення нового екземпляра з тими самими

властивостями на доступних ознаках. Аналізуючи по одному об'єкту за раз, SMOTE враховує різницю між спостереженням та його найближчим сусідом. Він збільшує різницю на випадкове число від нуля до одиниці. Потім визначає нову точку, додаючи випадкове число до об'єкта. Таким чином, SMOTE не копіює спостереження, а натомість створює новий, синтетичний на основі наявних даних [26]. На рисунку 3.5 приведено принцип роботи методу SMOTE.

```
# Divide by class
class_0 = train[train['fraud_reported'] == 0]
class_1 = train[train['fraud_reported'] == 1]
```

```
class_1_over = class_1.sample(count_class_0, replace=True)
test_over = pd.concat([class_0, class_1_over], axis=0)
print('Random over-sampling:')
print(test_over.fraud_reported.value_counts())
test_over.fraud_reported.value_counts().plot(kind='bar', title='Count (fraud_reported)');
```

```
Random over-sampling:
1    525
0    525
Name: fraud_reported, dtype: int64
```

Рисунок 3.4 – Балансування вибірки методом рандомного оверсемплінгу

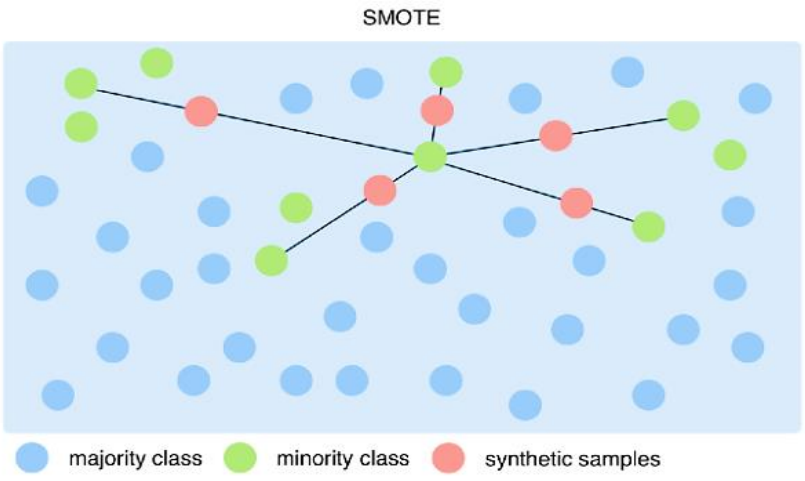

Рисунок 3.5 – Принцип роботи методу оверсемплінгу SMOTE

Реалізуємо оверсемплинг, використовуючи бібліотеку imbalanced-learn (imblearn), що створена для боротьби з проблемами незбалансованих наборів даних. Результат представлено на рисунку 3.6. З наведеної візуалізації можна побачити, що було створено 450 синтетичних шахрайських випадків, а класи відтепер рівно збалансовані 1:1.

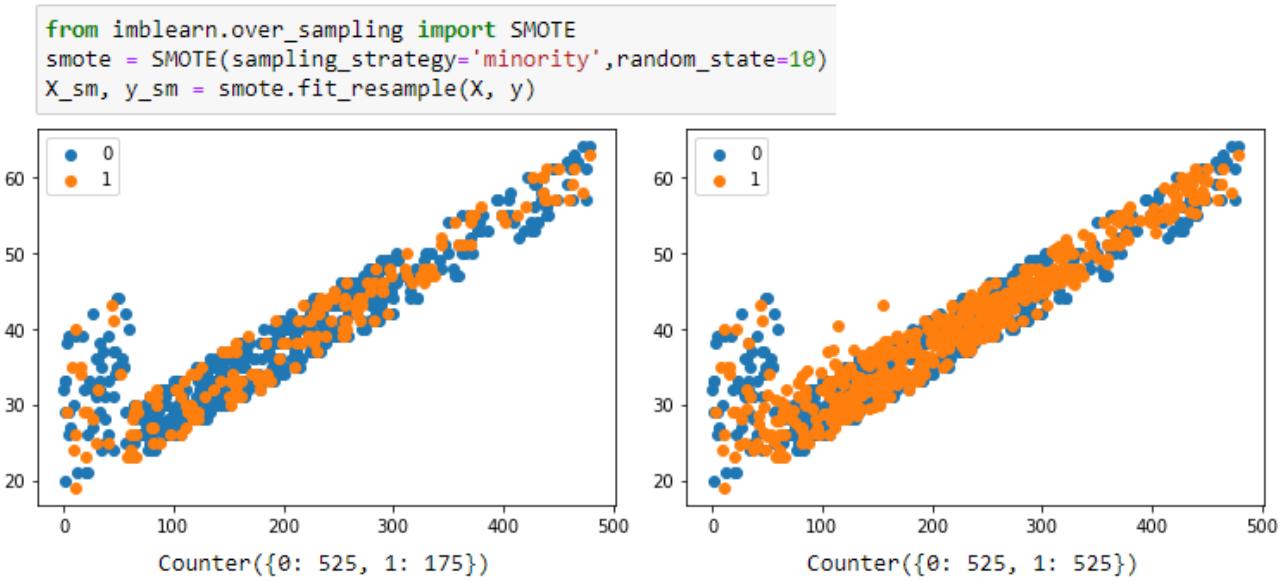

Рисунок 3.6 – Результат перебалансування методом SMOTE

Для спрощення згадки у тексті роботи отриманих наборів даних визначимо їх наступним чином:

– вибірка № 0: дані розділені на набір для навчання та тестування (70/30), співвідношення класів для обох підвибірок складає 75% випадків з відсутністю шахрайства на 25% шахрайських претензій;

– вибірка № 1: створена шляхом перебалансування класів рандомним оверсемплінгом таким чином, що навчальна вибірка збільшилась на 450 шахрайських випадків завдяки дублюванню існуючих записів;

– вибірка № 2: дані навчального набору були перебалансовані шляхом синтетичного створення для класу меншості 450 випадків наближених до шахрайських.

Усі ці три варіанти будуть використані для побудови моделі логістичної регресії на усіх змінних. Але перед тим, як подавати ці набори на навчання, необхідно виконати останній крок – шкалювання даних. Виконання цього кроку обумовлене тим, що дані за різними факторами змінюються в різних діапазонах. Зокрема, змінна «months_as_customer» змінюється від 0 до 479, в той час як «umbrella_limit» – від -1 000 000 до 7 993 220. При таких умовах помилки, зумовлені впливом другої змінної, будуть мати сильніший вплив на навчання, ніж першої. Шкалювання забезпечить рівний вплив кожної змінної [23].

Шкалювання екзогенних змінних будемо виконувати методом стандартизації, використовуючи клас StandardScaler. Ендогенна змінна не потребує шкалювання, вона є дискретною та приймає значення 0 або 1.

3.2 Побудова базової моделі

3.2.1 Логістична регресія на усіх змінних

Логістична регресія (Logistic regression) – це алгоритм побудови моделей бінарної класифікації та імовірнісного передбачення, який допомагає оцінити ймовірність того, що подія настане для конкретного об'єкта. Для оцінки параметрів рівняння логістичної регресії прийнято використовувати метод максимальної правдоподібності. Він полягає у підборі параметрів таким чином, щоб після підстановки даних у модель забезпечувалася максимальна ймовірність здійснення події, яка відповідає даним [30]. Логістична регресія є розширенням моделі лінійної регресії до завдань класифікації.

Перед початком моделювання для залежної змінної визначимо, що позитивною (positive) подією є шахрайський випадок, а випадок, де шахрайство відсутнє – негативною (negative).

Результатами моделювання будуть чотири варіанти класифікації:

– *TP* (True Positives) – істинно позитивні випадки, правильно класифіковані позитивні приклади;

– *TN* (True Negatives) – істинно негативні випадки, правильно класифіковані негативні приклади;

– *FN* (False Negatives) – хибнонегативні випадки, позитивні приклади, класифіковані як негативні;

– *FP* (False Positives) – хибнопозитивні випадки, негативні приклади, класифіковані як позитивні [23].

На основі отриманих результатів класифікації формується таблиця спряженості (confusion matrix – матриця неточностей), яка слугує для оцінки якості моделі (рисунок 3.7).

		Модель	
		Негативно	Позитивно
Фактично	NO	TN	FP
	YES	FN	TP

Рисунок 3.7 – Таблиця спряженості (Confusion Matrix)

Для аналізу результатів моделювання будуть використовуватись такі оцінки якості, як точність моделі (частка правильних прогнозів алгоритму):

$$AccuracyRate = \frac{TP + TN}{Total}, \tag{3.1}$$

та частка помилок:

$$ErrorRate = \frac{FP + FN}{Total}. \tag{3.2}$$

Проблема полягає в тому, що моделі, навчені на незбалансованих наборах даних, часто дають погані результати, коли їх намагаються узагальнити (щоб передбачати клас або класифікувати невидимі

спостереження). Незважаючи на обраний алгоритм, деякі моделі будуть більш сприйнятливими до незбалансованих даних, ніж інші. Хороша модель може не вийти через те, що алгоритм отримує значно більше прикладів одного класу, що спонукає його схилятися до нього. Він не дізнається, що робить інший клас «іншим», і не розуміє шаблонів, що лежать в основі і дозволяють розрізняти класи.

Алгоритм дізнається, що цей клас найпоширеніший, що робить його схильним до перенавчання класу більшості. Просто передбачивши клас більшості, моделі отримають високі оцінки. У таких випадках виникає «парадокс точності». Це відбувається при використанні метрики «точність» для виявлення найкращої моделі. А насправді ми можемо не мати хорошої моделі; а отримати модель, яка надає той самий клас всім спостереженням і погано узагальнює [26].

Щоб уникнути цієї проблеми, для оцінки якості будемо спиратись на такі показники, як чутливість (частка істинно позитивних випадків, що були правильно ідентифіковані моделлю):

$$Sensitivity = \frac{TP}{TP + FN}, \tag{3.3}$$

та специфічність (частка істинно негативних випадків, які були правильно ідентифіковані моделлю):

$$Specificity = \frac{TN}{TN + FP}. \tag{3.4}$$

Модель з високою чутливістю часто виявляє позитивні приклади, а модель з високою специфічністю – негативні [23]. Ними визначається об'єктивна цінність будь-якого бінарного класифікатора.

ROC-крива (Receiver Operator Characteristic curve) – це крива, яка найчастіше використовується для представлення результатів бінарної класифікації в машинному навчанні. ROC-крива показує залежність

кількості правильно класифікованих позитивних прикладів (True Positive Rate) від кількості неправильно класифікованих негативних прикладів (False Positive Rate). Дана крива являє собою лінію від (0,0) до (1,1) у координатах (TPR) і (FPR) [30], [31].

Одним із способів оцінити модель загалом, не прив'язуючись до конкретного порога, є AUC (Area Under Curve) – площа під кривою. У випадку, коли класифікатор не помиляється ($FPR = 0, TPR = 1$) площа під кривою дорівнюватиме одиниці. Коли ймовірності класів випадкова, AUC буде прагнути до 0,5 (класифікатор видаватиме однакову кількість TP та FP). Кожна точка графіку відповідає вибору деякого порога, а в ідеалі крива має прагнути точки (0,1). Можна вважати, що чим більший показник AUC, тим найкращою прогностичною силою наділена модель [30], [31].

Побудуємо модель логістичної регресії на навчальних наборах даних трьох вибірок, використовуючи усі змінні. Зробимо прогноз на тестових підвибірках та оцінимо результати класифікації, представлені у таблицях 3.1 – 3.3.

Таблиця 3.1 – Оцінка якості моделі логістичної регресії на усіх змінних вибірки № 0

Таблиця спряженості		Точність	0,89
212	**16**	Частка помилок	0,11
		Чутливість	0,76
17	**55**	Специфічність	0,92
		AUC	0,937

Таблиця 3.2 – Оцінка якості моделі логістичної регресії на усіх змінних вибірки № 1

Таблиця спряженості		Точність	0,87
196	**32**	Частка помилок	0,13
		Чутливість	0,903
7	**65**	Специфічність	0,859
		AUC	0,941

Таблиця 3.3 – Оцінка якості моделі логістичної регресії на усіх змінних вибірки № 2

Таблиця спряженості		Точність	0,866
195	33	Частка помилок	0,13
		Чутливість	0,902
7	65	Специфічність	0,855
		AUC	0,941

Отже, оцінимо результати першої моделі, яка буде слугувати орієнтиром. Бачимо, що модель правильно визначила 55 позитивних випадків (шахрайських) та 212 негативних (де шахрайство не встановлено). Однак у 11% випадків класифікатор помиляється: 17 шахрайських претензій були визначені як нешахрайські, тобто ті, що не несуть загрози, а 16 звичайних прикладів було розпізнано як шахрайські.

Хибнонегативні приклади (помилка 2 типу) становить основну загрозу, через те, що подія яка нас цікавить (шахрайство) помилково не виявляється, а класифікується як безпечна.

Показник специфічності значно вищій, ніж чутливості, це вказує на те, що модель орієнтована на визначення негативних випадків. Показник AUC доволі високий, тому загалом модель непогано визначає приналежність до класів. Класифікатор, що був навчений на даних вибірки № 0 орієнтується на виявленні негативних випадків, у той час, коли нас цікавлять саме позитивні.

Можна побачити, що варіанти моделі, які використовують додаткові створені випадки, показують вищі значення чутливості, тому краще базової виявляють позитивні шахрайські випадки. Отже, можна стверджувати, що реалізовані методи балансування виконують поставлену задачу та дають можливість покращити якість класифікації.

На даних вибірок № 1 та №2 показник специфічності трохи знизився, у порівнянні з базовою версію, але все одно має достатній рівень, а загальна прогностична сила усіх моделей має високе значення.

3.2.2 Логістична регресія на значущих змінних

Побудуємо моделі логістичної регресії на значущих змінних. Надалі будуть приведені результати моделювання на 1 та 2 вибірках, тому що для них характерна тенденція до виявлення позитивних випадків, тобто шахрайських претензій, які нас цікавлять найбільше, у той час коли моделі, засновані на даних вибірки 0, показують невисоку чутливість.

Виведемо статистичні показники моделі логістичної регресії на основі даних вибірки 1, які представлені на рисунку 3.8.

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
x1	0.1823	0.2551	0.7147	0.4748	-0.3177	0.6824
x2	-0.2573	0.2590	-0.9933	0.3206	-0.7650	0.2504
x3	0.0752	0.0966	0.7791	0.4359	-0.1141	0.2646
x4	0.1441	0.0992	1.4525	0.1464	-0.0503	0.3384
x5	-0.0372	0.0967	-0.3850	0.7002	-0.2267	0.1523
x6	-0.0762	0.1003	-0.7595	0.4475	-0.2728	0.1204
x7	0.0946	0.0974	0.9719	0.3311	-0.0962	0.2855
x8	0.0195	0.0951	0.2053	0.8373	-0.1669	0.2059
x9	0.0187	0.0976	0.1914	0.8482	-0.1725	0.2099
x10	0.2635	0.1007	2.6151	0.0089	0.0660	0.4609
x11	1.2853	0.1151	11.1629	0.0000	1.0596	1.5110
x12	0.2565	0.0980	2.6179	0.0088	0.0645	0.4486
x13	-0.1192	0.0970	-1.2279	0.2195	-0.3094	0.0710
x14	-0.2080	0.0964	-2.1570	0.0310	-0.3971	-0.0190
x15	-1.0056	0.3505	-2.8691	0.0041	-1.6926	-0.3187
x16	0.9282	0.2993	3.1016	0.0019	0.3416	1.5147
x17	1.6126	0.1088	14.8162	0.0000	1.3992	1.8259
x18	0.3018	0.1408	2.1437	0.0321	0.0259	0.5776
x19	0.2441	0.1027	2.3764	0.0175	0.0428	0.4454
x20	0.1513	0.0974	1.5523	0.1206	-0.0397	0.3422
x21	-0.0549	0.0997	-0.5506	0.5819	-0.2502	0.1404
x22	-0.3133	0.1099	-2.8512	0.0044	-0.5286	-0.0979
x23	0.0079	0.0960	0.0820	0.9347	-0.1803	0.1961
x24	0.0383	0.0943	0.4060	0.6848	-0.1465	0.2230
x25	0.2137	0.0950	2.2494	0.0245	0.0275	0.3999
x26	0.0164	0.0964	0.1704	0.8647	-0.1725	0.2053
x27	-114.7136	1159.8890	-0.0989	0.9212	-2388.0542	2158.6270
x28	21.6750	220.1967	0.0984	0.9216	-409.9025	453.2525
x29	22.7365	227.8532	0.0998	0.9205	-423.8476	469.3207
x30	81.7128	826.4568	0.0989	0.9212	-1538.1128	1701.5384
x31	-0.2609	0.1128	-2.3126	0.0207	-0.4821	-0.0398
x32	0.4730	0.1197	3.9531	0.0001	0.2385	0.7076
x33	-0.0962	0.0973	-0.9883	0.3230	-0.2869	0.0946

Рисунок 3.8 – Статистичні показники для даних вибірки 1

Критерієм відбору значущих змінних буде значення показника p-value у 3%. Змінні, за якими він буде перевищувати це значення, не будуть використані для побудови базової моделі. Отже, значущими змінними є

«insured_occupation», «insured_hobbies», «insured_relationship», «capital-loss», «incident_type», «collision_type», «incident_severity», «incident_state», «number_of_vehicles_involved», «witnesses», «auto_make», «auto_model», «authorities_contacted».

Проведемо навчання моделі на цих показниках та виведемо результати у таблиці 3.4.

Таблиця 3.4 – Оцінка якості моделі логістичної регресії на значущих змінних вибірки № 1

Таблиця спряженості		Точність	0,87
195	**33**	Частка помилок	0,13
		Чутливість	0,916
6	**66**	Специфічність	0,855
		AUC	0,939

Якість цієї моделі дещо зросла у порівнянні з моделлю від усіх змінних. Даний варіант буде вважатися базовим для цієї вибірки.

Виведемо статистичні показники для даних вибірки 2 (рисунок 3.9). Показник p-value призначимо рівним 3%, значущими змінними є «policy_csl», «insured_occupation», «insured_hobbies», «insured_relationship», «number_of_vehicles_involved», «collision_type», «incident_severity», «incident_state», «witnesses», «auto_model».

Навчимо модель на цих даних та виведемо результат у таблицю 3.5.

Таблиця 3.5 – Оцінка якості моделі логістичної регресії на значущих змінних вибірки № 2

Таблиця спряженості		Точність	0,88
197	**31**	Частка помилок	0,12
		Чутливість	0,93
5	**67**	Специфічність	0,864
		AUC	0,945

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
x1	0.2924	0.2642	1.1066	0.2685	-0.2255	0.8102
x2	-0.3254	0.2646	-1.2297	0.2188	-0.8440	0.1932
x3	0.0404	0.0987	0.4087	0.6828	-0.1532	0.2339
x4	0.2440	0.1029	2.3703	0.0178	0.0422	0.4458
x5	-0.0180	0.0995	-0.1811	0.8563	-0.2130	0.1770
x6	-0.0935	0.1003	-0.9320	0.3513	-0.2902	0.1031
x7	0.0195	0.1028	0.1896	0.8496	-0.1820	0.2210
x8	0.0999	0.0980	1.0194	0.3080	-0.0922	0.2920
x9	-0.0250	0.0990	-0.2525	0.8006	-0.2190	0.1690
x10	0.3319	0.1022	3.2464	0.0012	0.1315	0.5322
x11	1.3711	0.1201	11.4196	0.0000	1.1358	1.6064
x12	0.2961	0.0982	3.0149	0.0026	0.1036	0.4886
x13	-0.1166	0.0985	-1.1841	0.2364	-0.3096	0.0764
x14	-0.0825	0.0975	-0.8455	0.3978	-0.2737	0.1087
x15	-0.7073	0.3635	-1.9458	0.0517	-1.4197	0.0051
x16	0.7612	0.3178	2.3951	0.0166	0.1383	1.3842
x17	1.5713	0.1101	14.2769	0.0000	1.3556	1.7870
x18	0.2343	0.1473	1.5906	0.1117	-0.0544	0.5231
x19	0.2240	0.1033	2.1678	0.0302	0.0215	0.4266
x20	0.1404	0.0973	1.4428	0.1491	-0.0503	0.3312
x21	-0.0918	0.1014	-0.9053	0.3653	-0.2906	0.1070
x22	-0.2603	0.1121	-2.3213	0.0203	-0.4800	-0.0405
x23	-0.0074	0.0981	-0.0753	0.9399	-0.1997	0.1849
x24	0.0551	0.0970	0.5682	0.5699	-0.1350	0.2452
x25	0.2253	0.0989	2.2776	0.0227	0.0314	0.4192
x26	0.0239	0.0985	0.2421	0.8087	-0.1693	0.2170
x27	-56.9357	278.6463	-0.2043	0.8381	-603.0724	489.2010
x28	10.1075	50.7022	0.1993	0.8420	-89.2670	109.4819
x29	11.0809	53.4107	0.2075	0.8356	-93.6021	115.7639
x30	40.7575	199.4715	0.2043	0.8381	-350.1996	431.7145
x31	-0.0829	0.1162	-0.7131	0.4758	-0.3107	0.1449
x32	0.4252	0.1226	3.4691	0.0005	0.1850	0.6655
x33	-0.1632	0.1007	-1.6203	0.1052	-0.3605	0.0342

Рисунок 3.9 – Статистичні показники для даних вибірки 2

Якість даної моделі зросла у порівнянні з моделлю від усіх змінних, тому буде вважатися базовою для вибірки 2.

3.3 Побудова нелінійних класифікаторів

3.3.1 Метод опорних векторів

Машина опорних векторів (SVM (Support Vector Machine) є нелінійним узагальненням лінійного класифікатора, суть якого полягає у розширенні розмірності простору завдяки спеціальним ядерним функціям, що дозволяє виявити розділяючі гіперплощини між вихідними векторами значень [23]. Навчимо даний класифікатор на усіх змінних вибірки 1, використовуючи лінійну функцію, результат наведений у таблиці 3.6.

Таблиця 3.6 – Оцінка якості моделі опорних векторів на усіх змінних вибірки № 1

Таблиця спряженості		Точність	0,866
192	36	Частка помилок	0,13
		Чутливість	0,944
4	68	Специфічність	0,84
		AUC	0,932

Побудуємо класифікацію на значущих змінних вибірки №1, використовуючи лінійне ядро (таблиця 3.7).

Таблиця 3.7 – Оцінка якості моделі опорних векторів на значущих змінних вибірки № 1

Таблиця спряженості		Точність	0,866
192	36	Частка помилок	0,13
		Чутливість	0,944
4	68	Специфічність	0,84
		AUC	0,936

Можна бачити, що у цьому варіанті зріс показник AUC (у порівнянні з таблицею 3.6), проте ці моделі мають однакові результати класифікації.

Проведемо подібне навчання на даних вибірки 2, в обох випадках використовуючи лінійне ядро (таблиці 3.8 – 3.9).

Таблиця 3.8 – Оцінка якості моделі опорних векторів на усіх змінних вибірки № 2

Таблиця спряженості		Точність	0,863
192	36	Частка помилок	0,136
		Чутливість	0,931
5	67	Специфічність	0,84
		AUC	0,937

Таблиця 3.9 – Оцінка якості моделі опорних векторів на значущих змінних вибірки № 2

Таблиця спряженості		Точність	0,866
192	36	Частка помилок	0,13
		Чутливість	0,944
4	68	Специфічність	0,84
		AUC	0,94

Даний експеримент показує, що краще спрацювала модель, побудована на значущих змінних вибірки 2, про що каже вища точність визначення шахрайства та висока загальна точність моделі.

3.3.2 Метод k-найближчих сусідів

Метод k-NN (k-Nearest Neighbors) відноситься до непараметричних моделей, що здійснюють навчання на основі зразків. Такі моделі характеризуються запам'ятовуванням навчального набору даних.

Під час навчання алгоритм просто запам'ятовує вектори ознак спостережень та його мітки класів (зразки). Задається параметр алгоритму k, який визначає кількість «сусідів», які будуть використовуватися при класифікації. На етапі класифікації для нового об'єкта визначається k найближчих попередньо класифікованих спостережень. Обирається клас, якому належить більшість з k найближчих прикладів-сусідів, і об'єкт відносять до цього класу [32].

Алгоритм k-NN можна представити у вигляді наступних кроків:

– вибрати число k та метрику відстані.

– знайти k найближчих сусідів зразка, який потрібно класифікувати.

– призначити мітку класу з більшості голосів [33].

Алгоритм k-NN є чутливим до дисбалансу класів у навчальних даних, він «схильний» до зміщення рішення у бік домінуючого класу, оскільки такі об'єкти просто частіше потрапляють до числа найближчих сусідів.

Визначимо оптимальне значення параметру k за даними вибірки 1 на усіх змінних (рисунок 3.10).

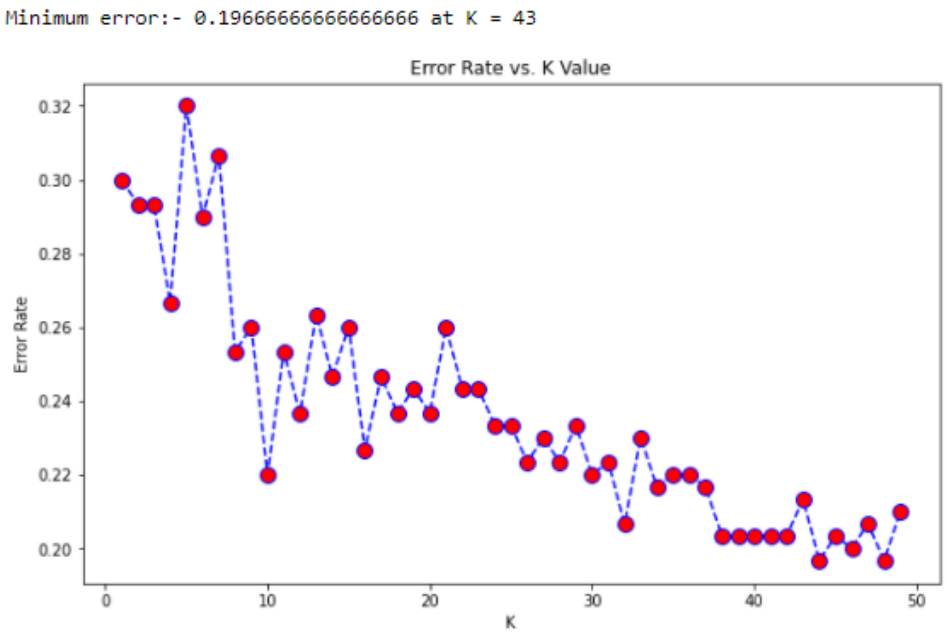

Рисунок 3.10 – Визначення оптимального значення k для даних вибірки 1

Мінімальне значення помилки досягається при $k = 43$. Навчимо модель, використовуючи 43 сусіда, результати представимо у таблиці 3.10.

Таблиця 3.10 – Оцінка якості моделі k-NN на усіх змінних вибірки № 1

Таблиця спряженості		Точність	0,787
180	**48**	Частка помилок	0,21
		Чутливість	0,777
16	**56**	Специфічність	0,79
		AUC	0,84

Модель, побудована на такому наборі даних, погано визначає як позитивні, так і негативні приклади. Також дана модель поки що має найнижче значення специфічності серед усіх та велику кількість хибнонегативних випадків. Експеримент зі зміною значення k параметру суттєво не впливає на результат класифікації.

Здійснимо моделювання на значущих змінних, визначивши оптимальним $k = 15$; результати наведені у таблиці 3.11.

Таблиця 3.11 – Оцінка якості моделі k-NN на значущих змінних вибірки № 1

Таблиця спряженості		Точність	0,857
189	**39**	Частка помилок	0,14
		Чутливість	0,944
4	**68**	Специфічність	0,82
		AUC	0,889

Класифікація на значущих змінних дає високе значення чутливості, модель добре визначає шахрайські випадки, але при цьому гірше інших розпізнає приклади, де шахрайство відсутнє.

Перевіримо результат роботи алгоритму k-NN на даних вибірки 2. Найменше значення помилки для моделі на усіх змінних досягається при $k = 1$. Тому навчимо модель, використовуючи одного сусіда. Результати наведені у таблиці 3.12.

Таблиця 3.12 – Оцінка якості моделі k-NN на усіх змінних вибірки № 2

Таблиця спряженості		Точність	0,63
137	**91**	Частка помилок	0,37
		Чутливість	0,7
21	**51**	Специфічність	0,6
		AUC	0,843

Класифікатор, побудований на цих даних, помиляється у 37% випадків. Велика кількість неправильно класифікованих прикладів підтверджується низькими значеннями показників якості моделювання. Кількість хибнонегативних випадків велика.

Подивимось, як буде поводитись даний класифікатор на значущих змінних. Навчимо модель на 17 сусідах, результат приведемо у таблиці 3.13.

Таблиця 3.13 – Оцінка якості моделі k-NN на значущих змінних вибірки №2

Таблиця спряженості		Точність	0,84
185	**43**	Частка помилок	0,16
		Чутливість	0,931
5	**67**	Специфічність	0,811
		AUC	0,906

Модель значно краще визначає шахрайство на значущих фактрах, про що каже високий показник чутливості. При цьому специфічність дещо знижена у порівнянні з іншими моделями.

3.3.3 Баєсова класифікація

Наївний байєсовський класифікатор (Naive Bayes Classifier) – простий ймовірнісний класифікатор, який використовує теорему Байєса для визначення ймовірності приналежності об’єкта до одного з класів, з припущенням про незалежність подій. Ймовірність події може бути скоригована в міру введення нових даних [23].

Оцінимо роботу даного класифікатора на даних вибірки 1, результати представлені у таблицях 3.14 – 3.15. Можемо побачити, що даний класифікатор значно краще працює на значущих змінних, однак в обох випадках рівень визначення негативних випадків невисокий.

Таблиця 3.14 – Оцінка якості моделі Баєсової класифікації на усіх змінних вибірки № 1

Таблиця спряженості		Точність	0,62
123	**105**	Частка помилок	0,37
		Чутливість	0,88
8	**64**	Специфічність	0,54
		AUC	0,852

Таблиця 3.15 – Оцінка якості моделі Баєсової класифікації на значущих змінних вибірки №1

Таблиця спряженості		Точність	0,73
156	72	Частка помилок	0,27
		Чутливість	0,875
9	63	Специфічність	0,68
		AUC	0,873

Перевіримо, які результати дасть використання цього методу на даних вибірки 2, результати наведені у таблицях 3.16 – 3.17.

Таблиця 3.16 – Оцінка якості моделі Баєсової класифікації на усіх змінних вибірки № 2

Таблиця спряженості		Точність	0,646
131	97	Частка помилок	0,35
		Чутливість	0,875
9	63	Специфічність	0,57
		AUC	0,85

Таблиця 3.17 – Оцінка якості моделі Баєсової класифікації на значущих змінних вибірки № 2

Таблиця спряженості		Точність	0,86
192	36	Частка помилок	0,14
		Чутливість	0,916
6	66	Специфічність	0,84
		AUC	0,926

На даних вибірки 2 класифікатор також краще визначає шахрайські випадки при моделюванні на значущих змінних. Алгоритм, побудований від усіх змінних нараховує велику кількість хибнопозитивних визначень, а як наслідок низький показник специфічності при довільному рівні чутливості.

3.3.4 Дерево рішень

Класифікатори на основі дерева рішень (Decision Tree) розбиває дані, задаючи послідовність питань. Базуючись на ознаках у навчальному наборі, модель на основі дерева рішень навчається послідовності питань, щоб виводити мітки класів для зразків. Використовуючи алгоритм прийняття рішення, ми починаємо з кореня дерева і розбиваємо дані за ознакою, що дає в результаті найбільший приріст інформації.

В рамках ітераційного процесу описана процедура розбиття повторюється в кожному вузлі, поки листові вузли не стануть чистими. Це означає, що всі зразки в кожному вузлі належать тому самому класу. На практиці результатом може бути дуже глибоке дерево з численними вузлами, що легко здатне призвести до перенавчання. Щоб уникнути такої проблеми дерево зазвичай підрізається, таким чином встановлюється межа його максимальної глибини [33].

Навчимо класифікатор на основі дерева рішень на даних вибірки 1, використовуючи параметр максимальної глибини листів $depth = 4$. Результати наведені у таблиці 3.18.

Таблиця 3.18 – Оцінка якості моделі дерева рішень на усіх змінних вибірки № 1

Таблиця спряженості		Точність	0,866
192	**36**	Частка помилок	0,13
		Чутливість	0,944
4	**68**	Специфічність	0,84
		AUC	0,913

Даний алгоритм дає високе значення чутливості та разом з цим доволі високе значення специфічності. Кількість помилок 2 типу низька, а загальна точність та прогностична сила класифікатора високі. Модель, яка навчалась на значущих змінних, видає такий самий результат.

Перевіримо роботи алгоритму на даних вибірки 2 (з $depth = 4$). Результати приведені у таблиці 3.19.

Таблиця 3.19 – Оцінка якості моделі дерева рішень на усіх змінних вибірки № 2

Таблиця спряженості		Точність	0,866
192	**36**	Частка помилок	0,13
		Чутливість	0,944
4	**68**	Специфічність	0,84
		AUC	0,907

Побудована модель теж демонструє одні з найкращих показників якості, алгоритм добре розпізнає як позитивні, так і негативні випадки. Значення чутливості та AUC підтверджують відповідність класифікатора заданим вимогам.

Класифікатори на основі дерева рішень однаково працюють на двох вибірках на усіх наборах змінних.

3.3.5 Випадковий ліс

Випадковий ліс (Random Forest) можна розглядати як ансамбль дерев рішень. В основі випадкового лісу закладена ідея усереднення множини (глибоких) дерев прийняття рішень, які окремо страждають від високої дисперсії, з метою побудови більш надійної моделі, що має більшу ефективність узагальнення та меншу сприйнятливість до перенавчання [33].

Навчимо модель випадкового лісу на даних вибірки 1. Для моделі на усіх змінних встановимо параметр кількості дерев рівний 8, глибини – 5, для моделі на значущих факторах 5 та 5 відповідно, показники якості наведені у таблицях 3.20 – 3.21.

Таблиця 3.20 – Оцінка якості моделі випадкового лісу на усіх змінних вибірки №1

Таблиця спряженості		Точність	0,846
194	34	Частка помилок	0,15
		Чутливість	0,83
12	60	Специфічність	0,85
		AUC	0,883

Таблиця 3.21 – Оцінка якості моделі випадкового лісу на значущих змінних вибірки №1

Таблиця спряженості		Точність	0,873
195	33	Частка помилок	0,126
		Чутливість	0,93
5	67	Специфічність	0,855
		AUC	0,914

Якщо порівнювати ці дві моделі, то класифікатор, побудований на значущих факторах дає значно більше значення чутливості та AUC. Моделі на основі випадкового лісу гірше виконують завдання класифікації, ніж алгоритм дерева рішень.

Перевіримо, як працює даний алгоритм на даних вибірки 2. Для обох моделей встановимо параметр кількості дерев рівний 15, глибини – 8, для моделі на значущих факторах – аналогічні значення. Результати наведені у таблицях 3.22 – 3.23.

Таблиця 3.22 – Оцінка якості моделі випадкового лісу на усіх змінних вибірки №2

Таблиця спряженості		Точність	0,85
205	23	Частка помилок	0,146
		Чутливість	0,71
21	51	Специфічність	0,899
		AUC	0,909

Класифікатор, що побудовано на усіх змінних, показує кращі результати виявлення нешахрайських претензій, це доводить висока специфічність. При цьому фіксується значна кількість хибнонегативних випадків.

Таблиця 3.23 – Оцінка якості моделі випадкового лісу на значущих змінних вибірки №2

Таблиця спряженості		Точність	0,87
198	**30**	Частка помилок	0,13
		Чутливість	0,875
9	**63**	Специфічність	0,868
		AUC	0,922

Класифікатор, що побудовано на усіх змінних, показує кращі результати виявлення нешахрайських випадків, це доводить достатня специфічність при достатньому рівні чутливості. На даних вибірки 2 алгоритм випадкового лісу також показує гірші результати роботи, ніж метод дерева рішень.

3.3.6 Класифікаційна нейронна мережа (CNN)

Побудуємо класифікатор на основі нейронної мережі прямої передачі сигналу. Так як ми маємо справу з моделлю класифікації, то необхідно створювати архітектуру, що звужує: кількість нейронів на першому шарі не повинна перевищувати кількість нейронів на вхідному шарі.

Створимо нейронну мережу на даних вибірки 1. На початку включимо в модель всі змінні, тому на вхідному шарі кількість нейронів дорівнюватиме 33, а для першого шару візьмемо кількість, що дорівнює вхідному. Використовуємо функцію активації «linear». Для другого шару залишимо 1 нейрон, тому що ми маємо всього два класи (кількість виходів у нашому випадку = 2-1). Будемо використовувати сигмоїдальну функцію активації, оскільки вона змінюється в діапазоні від 0 до 1, що відповідає

нашій ендогенній змінній. Варто зазначити, що такі налаштування нейронної мережі були остаточно взяті після низки підборів значень та змін точності моделі. Навчимо модель на встановлених параметрах, результати наведемо у таблиці 3.24.

Таблиця 3.24 – Оцінка якості моделі CNN на усіх змінних вибірки №1

Таблиця спряженості		Точність	0,863
192	36	Частка помилок	0,136
		Чутливість	0,931
5	67	Специфічність	0,84
		AUC	0,94

Побудуємо нейронну мережу на значущих змінних, їх кількість дорівнює 13, тому на вхідному шарі матимемо 13 нейронів, а на першому шарі візьмемо кількість рівну вхідному. Результати роботи класифікатора приведені у таблиці 3.25.

Таблиця 3.25 – Оцінка якості моделі CNN на значущих змінних вибірки №1

Таблиця спряженості		Точність	0,89
201	27	Частка помилок	0,11
		Чутливість	0,916
6	66	Специфічність	0,88
		AUC	0,94

Порівняємо дві отримані моделі: показники чутливості та AUC мають високе значення в обох випадках, але чутливість вища при включенні лише значущих змінних. Кількість істинно позитивних випадків більша, а хибнонегативних менша.

Побудуємо нейронні мережі для даних вибірки 2. Для моделі на усіх змінних архітектура буде наступною: 33-33-1; на значущих змінних: 10-9-1. Результати класифікації наведемо у таблицях 3.26 – 3.27.

Таблиця 3.26 – Оцінка якості моделі CNN на усіх змінних вибірки №2

Таблиця спряженості		Точність	0,876
198	**30**	Частка помилок	0,123
		Чутливість	0,902
7	**65**	Специфічність	0,868
		AUC	0,941

Таблиця 3.27 – Оцінка якості моделі CNN на значущих змінних вибірки №2

Таблиця спряженості		Точність	0,88
197	**31**	Частка помилок	0,12
		Чутливість	0,931
5	**67**	Специфічність	0,864
		AUC	0,945

На даних цієї вибірки краще спрацювала модель, побудована на значущих змінних, бо спостерігається більша кількість правильно класифікованих випадків шахрайства. Моделі на основі нейронної мережі дають можливість одночасно добре визначати як негативні показники, так і позитивні, значення специфічності та AUC є одними з найвищих для цього набору даних.

3.4 Висновки до розділу 3

Отримавши результати класифікації за усіма методами на двох створених вибірках, проведемо порівняльний аналіз якості моделювання. Помістимо у таблицю табл. 3.28 оціночні показники з усіх проведених експериментів, де позначено AC – AccuracyRate (3.1), ER – ErrorRate (3.2), SE – Sensitivity (3.3). SP – Specificity (3.4).

Таблиця 3.28 – Оцінки якості моделей класифікації за усіма наборами даних

Модель	Вибірка №1					Вибірка №2				
	AC	**ER**	**SE**	**SP**	***AUC***	**AC**	**ER**	**SE**	**SP**	***AUC***
LR (усі змінні)	0,870	0,130	0,903	0,859	0,941	0,866	0,130	0,902	0,855	0,941
значущі	0,870	0,130	0,916	0,855	0,939	0,880	0,120	0,930	0,864	0,945
SVM (усі змінні)	0,866	0,130	0,944	0,840	0,932	0,863	0,136	0,931	0,840	0,937
значущі	0,866	0,130	0,944	0,840	0,936	0,866	0,130	0,944	0,840	0,940
KNN (усі змінні)	0,787	0,210	0,770	0,790	0,840	0,630	0,370	0,700	0,600	0,843
значущі	0,857	0,140	0,944	0,820	0,889	0,840	0,160	0,931	0,811	0,906
NB (усі змінні)	0,620	0,370	0,880	0,540	0,852	0,646	0,350	0,875	0,570	0,850
значущі	0,730	0,270	0,875	0,680	0,873	0,860	0,140	0,916	0,840	0,926
DT (усі змінні)	0,866	0,130	0,944	0,840	0,913	0,866	0,130	0,944	0,840	0,907
RF (усі змінні)	0,846	0,150	0,830	0,850	0,883	0,850	0,146	0,710	0,899	0,909
значущі	0,873	0,126	0,930	0,855	0,914	0,870	0,130	0,875	0,868	0,922
CNN (усі змінні)	0,863	0,136	0,931	0,840	0,940	0,876	0,123	0,902	0,868	0,941
значущі	0,890	0,110	0,916	0,880	0,940	0,880	0,120	0,931	0,864	0,945

Для даних вибірки 1 (нагадаємо, що дана вибірка була створена шляхом перебалансування класів рандомним оверсемплінгом таким чином, щоб до навчального набору потрапило якомога більше випадків з встановленим фактом шахрайства) за всіма методами кращі результати класифікації показують моделі, що були побудовані на значущих факторах. Це вказує на правильність відбору показників, що впливають на виявлення шахрайства.

Найгірша якість класифікації спостерігається за методом Байєсовського класифікатора, а також k-найближчих сусідів та випадкового лісу, побудованих на усіх змінних. Найточніше класифікують шахрайські

випадки на цьому наборі даних логістична регресія, метод опорних векторів, дерево рішень, нейронна мережа та метод випадкового лісу на значущих факторах. Метод k-NN показує високу чутливість, у той час коли значення специфічності у порівнянні з іншими моделями знижене. Якщо ухвалювати рішення про використання певного методу, то я б відштовхувалась не тільки від високого показника чутливості, а й від співвідношення різних оцінок. У рамках даної задачі нам звісно важливіше виявити шахрайські випадки, але в той же час велике значення має відтворення реальної картини. Через це я буду орієнтуватись на високу частку визначення позитивних випадків у поєднанні з високим рівнем виявлення негативних випадків та загальну прогностичну силу моделі. Усі класифікатори, що визнані найкращими на даних цієї підходять під ці критерії.

Аналіз результатів якості моделей, побудованих на даних вибірки 2 дає розуміння, що використані методи, як і першому випадку, краще класифікують випадки при включенні усіх змінних в модель. Гірші результати, показують метод випадкового лісу, а також Байєсовський класифікатор та метод k-найближчих сусідів на усіх змінних. Високі показники якості показують моделі на значущих змінних, що були отримані на основі логістичної регресії, опорних векторів, дерева рішень, нейронної мережі та Байєсовського класифікатора. Метод k-найближчих сусідів, що побудовано на значущих змінних, добре визначає випадки шахрайства, однак має нижчу специфічність та загальну точність, бо погано класифікує негативні випадки. Найкращими моделями, створеними на основі даних вибірки 2, стали нейронна мережа, логістична регресія, дерево рішень та метод опорних векторів.

За результатами проведеного моделювання та аналізу якості можна стверджувати, що найкращим рішенням при виборі варіанту класифікатора для виявлення шахрайських претензій стануть методи логістичної регресії, дерева рішень, SVM та CNN.

Якщо порівнювати результати класифікації випадків за даними двох

вибірок, то вища якість спостерігається при виборі даних вибірки 2, де більшість шахрайських випадків є штучно створеними. Цей підхід ефективний, тому що створюються нові синтетичні приклади з класу меншості, які є правдоподібними, тобто відносно близькі по простору ознак до існуючих прикладів з класу меншості. Загальним недоліком цього підходу є те, що синтетичні приклади створюються без урахування класу більшості, що може привести до неоднозначних прикладів, якщо існує сильне перекриття класів [29].

В завданнях виявлення шахрайства чи кредитного скорингу нам потрібне більше ранжування, тобто визначення клієнтів більш схильних до шахрайства, ніж інші. В таких випадках баланс класів може бути не таким важливим, оскільки пороги для прийняття рішень все одно вибираються вручну, виходячи з економічних міркувань. Кориснішим може стати передбачення справжньої ймовірності шахрайства, тому семплювання, що буде спотворювати ці ймовірності, є небажаним.

Загальні результати проведеного аналізу страхових випадків з автострахування дають зрозуміти, що при розгляді претензій особливу увагу слід звертати на наступні показники:

– характеристику клієнта: потенційну загрозу становлять страхувальники, що займають посади топ-менеджерів, мають професіональну юридичну чи медичну освіту, не мають сім'ї чи класифіковані як «інший родич», а в якості хобі обирають шахи чи кросфіт;

– деталі інциденту: для виявлення шахрайських претензій найзначущою ознакою є тяжкість інциденту, яка визначається як велике пошкодження. Характерні два типи інциденту: зіткнення кількох авто та одиночне зіткненні. Для шахрайських випадків найпопулярнішим типом зіткнення є зіткнення ззаду. Найчастіше фіксується задіяння одного транспортного засобу та трьох. Інцидент у більшості випадків має двох свідків. У шахрайських претензіях найбільш популярні моделі автомобілів: «RAM», «A3», «F150», «Jetta».

ВИСНОВКИ

В рамках даної кваліфікаційної роботи магістра було проведено аналіз сучасного стану української страхової сфери та досліджені основні показники діяльності ринку. Результати дослідження показують, що рівень його розвитку можна оцінити як недостатній чи низький, він доволі сильно відстає від розвинених ринків європейських країн та США. Частка ринку страхування України у складі світового ринку дуже незначна, як низька і частка у ВВП країни. Відмічається недовіра населення до страхування, причиною цього є ненадійність страхових компаній, які схильні до банкрутства. Спостерігається тенденція до значного зменшення гравців на ринку. Найпоширенішими видами страхування є автострахування (КАСКО, ОСЦПВ та Зелена картка), та особисте страхування (життя та медичне). Найбільший коефіцієнт збитковості спостерігається за обов'язковим та добровільним автострахуванням та медичним. Основну частку усіх договорів займає страхування від нещасних випадків на транспорті, яке є обов'язковим.

Були вивчені способи застосування машинного навчання в страховому бізнесі, що покликані вирішувати різноманітні завдання, одним з яких є виявлення шахрайських випадків. Серед усіх випадків на українському ринку, за якими страхові компанії несуть збитки, майже 2% припадає на випадки з встановленим фактом шахрайства. Як у світі, так і в Україні основним видом страхування, де спостерігається найбільша кількість махінацій є автострахування. В роботі приводиться короткий опис вже існуючих робіт з виявлення шахрайства у автострахуванні. Зазначена тема є доволі поширеною серед зарубіжних дослідників, у той час коли для робіт вітчизняних науковців дана проблема не є популярною.

Для реалізації поставленої мети проведено аналіз бази даних з 1000 страхових претензій клієнтів, у 25% з яких зафіксовано факт шахрайства. З наявних 39 змінних для подальшої обробки залишено 34 найбільш

інформативні. Виконані етапи підготовки даних, необхідної для виконання подальшого моделювання, та вивчені статистичні показники. Одним із ключових результатів аналізу стало визначення найвпливовіших факторів, за якими проаналізовані статистики та побудовані візуалізації, що дозволяють наочно простежити залежності між показниками та їх відношення до результуючої змінної.

Таким чином було визначено, що потенційну загрозу становлять страхувальники, що займають посади топ-менеджерів, мають професіональну юридичну чи медичну освіту, не мають сім'ї чи класифіковані як «інший родич», а в якості хобі обирають шахи чи кросфіт. Також було встановлено, що при розгляді претензій особливу увагу слід звертати на наступні деталі інциденту: для шахрайських претензій найзначущою ознакою є тяжкість інциденту, яка визначається як велике пошкодження. Характерні два типи інциденту: зіткнення кількох авто та одиночне зіткненні; найпопулярнішим типом зіткнення є зіткнення ззаду. Найчастіше фіксується задіяння одного транспортного засобу та трьох. У шахрайських претензіях найбільш популярні моделі автомобілів: «RAM», «A3», «F150», «Jetta». Інцидент у більшості випадків має двох свідків.

Розглянуто вирішення проблеми незбалансованих даних та реалізовані методи боротьби з ними, які полягають у навчанні моделей на якомога більшій кількості шахрайських претензій та застосуванні техніки передискретизації. Дані були перебалансовані за використанням методу SMOTE, суть якого полягає у синтетичному створенні для класу меншості випадків, наближених до шахрайських.

У ході дослідження на отриманих наборах даних були побудовані класифікатори на основі логістичної регресії, методу опорних векторів, алгоритму k-найближчих сусідів, Байєсовського класифікатора, дерева рішень, випадкового лісу та нейронної мережі. За допомогою проведення порівняльного аналізу показників якості класифікації визначені найкращі методи для виявлення шахрайських претензій. Для обох вибірок такими

методами були визнані логістична регресія, метод опорних векторів, дерево рішень та класифікаційна нейронна мережа, які одночасно мають високу частку визначення позитивних випадків у поєднанні з високим рівнем виявлення негативних випадків та загальну прогностичну силу моделі. Оцінка показників якості побудованих класифікаторів вказує на можливість досягнення вищої точності класифікації при використанні даних, що були перебалансовані.

Отримані в ході дослідження моделі демонструють доволі високі результати класифікації, але набір даних, на яких проходило навчання, обмежений кількістю випадків. До того ж клас шахрайства, визначенню якого приділяється найбільша увага, складає усього чверть від усіх даних. Тому для покращення результатів моделювання, виявлення закономірностей утворення шахрайських випадків, точнішого визначення впливових факторів та потенційно небезпечних клієнтів необхідно проводити дослідження на даних більш об'ємної вибірки.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

1. Про страхування: Закон України від 07 бер. 1996 р. № 85/96-ВР. URL: https://zakon.rada.gov.ua/laws/show/85/96-%D0%B2%D1%80#Text (дата звернення: 10.05.2024).

2. Горбач Л. М., Кадебська Е. В. Страхування: підручник. Київ : Кондор Видавництво, 2016. 544 с.

3. Яцкевич І. В., Голинська. О. В. Фінанси: навч. посіб. Одеса : ОРІДУ НАДУ, 2014. 316 с.

4. Ruda O. Insurance market development in ukraine. *Efektyvna ekonomika*. 2020. No. 2. URL: https://doi.org/10.32702/2307-2105-2020.2.55 (date of access: 10.05.2024).

5. Статистика страхового ринку України. *FORINSURER: Форіншурер – журнал про страхування та InsurTech*. URL: https://forinsurer.com/stat (дата звернення: 10.05.2024).

6. Національний Банк України. URL: https://bank.gov.ua/ (дата звернення: 10.05.2024).

7. Золотарьова О.В. Ключові тенденції та пріоритети розвитку ринку страхових послуг в Україні. Економіка і суспільство. 2017. №11. С. 413–420.

8. Прокопчук Е. Т. Оценка современного состояния и перспективы функционирования страхового рынка Украины. *Финансы и кредит*. 2017. т. 23. Вып. 12. С. 709–730.

9. Лащик І., Кондрат І., Віблий П., Білець В. Страховий ринок України: сучасний стан та перспективи розвитку. *Галицький економічний вісник*. 2020. № 5 (66). С. 105–112.

10. Pratt M. K. 10 common uses for machine learning applications in business. *Enterprise AI*. URL: https://searchenterpriseai.techtarget.com/feature/10-common-uses-for-machine-learning-applications-in-business (date of access: 10.05.2024).

11. Top 10+ awesome machine learning applications in 2023. *ProjectPro*. URL: https://www.projectpro.io/article/10-awesome-machine-learning-applications-of-today/364 (date of access: 10.05.2024).

12. 6 ways machine learning is changing insurance [updated]. *Digital Acceleration Company | Netguru*. URL: https://www.netguru.com/blog/machine-learning-insurance (date of access: 10.05.2024).

13. Аферисти з фантазією: як клієнти страхових компаній намагаються облудним шляхом отримати виплати. *Mind.ua*. URL: https://mind.ua/publications/20204635-aferisti-z-fantazieyu-yak-klienti-strahovih-kompanij-namagayutsya-obludnim-shlyahom-otrimati-viplati (дата звернення: 10.05.2024).

14. Hassan A. K. I., Abraham A. Modeling insurance fraud detection using imbalanced data classification. *Advances in intelligent systems and computing*. Cham, 2015. P. 117–127. URL: https://doi.org/10.1007/978-3-319-27400-3_11 (date of access: 10.05.2024).

15. Performance comparative study of machine learning algorithms for automobile insurance fraud detection / B. Itri et al. *2019 third international conference on intelligent computing in data sciences (ICDS)*, Marrakech, Morocco, 28–30 October 2019. 2019. URL: https://doi.org/10.1109/icds47004.2019.8942277 (date of access: 10.05.2024).

16. Patel D. K., Subudhi S. Application of extreme learning machine in detecting auto insurance fraud. *2019 international conference on applied machine learning (ICAML)*, Bhubaneswar, India, 25–26 May 2019. 2019. URL: https://doi.org/10.1109/icaml48257.2019.00023 (date of access: 10.05.2024).

17. Pattanaik A., Panigrahi S. Use of particle swarm optimization for feature selection and data mining methods for efficient detection of automobile insurance fraud. *2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE)*, Bhubaneswar, India, 27–28 July 2018. 2018. URL:

https://doi.org/10.1109/icrieece44171.2018.9009411 (date of access: 10.05.2024).

18. Mubarek A. M., Adali E. Multilayer perceptron neural network technique for fraud detection. *2017 International Conference on Computer Science and Engineering (UBMK)*, Antalya, 5–8 October 2017. 2017. URL: https://doi.org/10.1109/ubmk.2017.8093417 (date of access: 10.05.2024).

19. Ghorbani A., Farzai S. Fraud detection in automobile insurance using a data mining based approach. *International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)*. 2018. Vol. 8(27). P. 3767–3771.

20. Hanafy M., Ming R. Using machine learning models to compare various resampling methods in predicting insurance fraud. *Journal of Theoretical and Applied Information Technology*. 2021. Vol. 99. № 12. P. 2819–2833.

21. Yan C., Li Y. The identification algorithm and model construction of automobile insurance fraud based on data mining. *2015 Fifth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC)*, Qinhuangdao, China, 18–20 September 2015. 2015. URL: https://doi.org/10.1109/imccc.2015.408 (date of access: 10.05.2024).

22. Research and application of random forest model in mining automobile insurance fraud / Y. Li et al. *2016 12th international conference on natural computation and 13th fuzzy systems and knowledge discovery (ICNC-FSKD)*, Changsha, China, 13–15 August 2016. 2016. URL: https://doi.org/10.1109/fskd.2016.7603443 (date of access: 10.05.2024).

23. Кононова К. Ю. Машинне навчання: методи та моделі: підручник для бакалаврів, магістрів та докторів філософії спеціальності 051 «Економіка». Харків: ХНУ імені В. Н. Каразіна, 2020. 301 с.

24. Roy B. All about categorical variable encoding. *Medium*. URL: https://towardsdatascience.com/all-about-categorical-variable-encoding-305f3361fd02 (date of access: 10.05.2024).

25. McKinney W. Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython, 2nd Edition. O'Reilly Media, 2017. 544 p.

Lahera G. Unbalanced datasets & what to do. *Medium*. URL: https://medium.com/strands-tech-corner/unbalanced-datasets-what-to-do-144e0552d9cd (date of access: 10.05.2024).

26) Agarwal R. The 5 sampling algorithms every data scientist need to know. *Medium*. URL: https://towardsdatascience.com/the-5-sampling-algorithms-every-data-scientist-need-to-know-43c7bc11d17c (date of access: 10.05.2024).

27) Resampling strategies for imbalanced datasets. *Kaggle: Your Machine Learning and Data Science Community*. URL: https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t3 (date of access: 10.05.2024).

28) SMOTE for imbalanced classification with python - machinelearningmastery.com. *MachineLearningMastery.com*. URL: https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/ (date of access: 10.05.2024).

29) Zou K. H., O'Malley A. J., Mauri L. Receiver-Operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*. 2007. Vol. 115, no. 5. P. 654–657. URL: https://doi.org/10.1161/circulationaha.105.594929 (date of access: 10.05.2024).

30) On the evaluation of unsupervised outlier detection: measures, datasets, and an empirical study / G. O. Campos et al. *Data mining and knowledge discovery*. 2016. Vol. 30, no. 4. P. 891–927. URL: https://doi.org/10.1007/s10618-015-0444-8 (date of access: 10.05.2024).

31) Hall P., Park B. U., Samworth R. J. Choice of neighbor order in nearest-neighbor classification. *The annals of statistics*. 2008. Vol. 36, no. 5. P. 2135–2152. URL: https://doi.org/10.1214/07-aos537 (date of access: 10.05.2024).

32) Raschka S., Mirjalili V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow2, 3rd Edition. Birmingham: Packt Publishing, 2019. 770 p.