УДК 519.2:368
DOI 10.33111/nfmte.2020.138

ВИЯВЛЕННЯ ШАХРАЙСТВА В АВТОСТРАХУВАННІ: ПРОБЛЕМА НЕЗБАЛАНСОВАНОЇ ВИБІРКИ

К. Ю. Кононова

Харківський національний університет імені В.Н. Каразіна
майдан Свободи, 4, м. Харків, 61022, Україна
ORCID: 0000-0001-6990-5746, E-mail: kateryna.kononova@karazin.ua

А. С. Гавриленко

Харківський національний університет імені В.Н. Каразіна
майдан Свободи, 4, м. Харків, 61022, Україна
ORCID: 0000-0002-9191-6238, E-mail: anna.gavrilenko99@gmail.com

Вирішуючи завдання класифікації методами машинного навчання, фахівці з аналізу даних часто стикаються з проблемою незбалансованих даних. Наявність дисбалансу класів характерна для даних фінансового сектору, зокрема для задач з виявлення шахрайства в автострахуванні. Навчання моделей на незбалансованих даних може призвести до неправильної класифікації та великої кількості помилкових визначень через схильність класифікатора відносити випадки до класу більшості.

Дана робота присвячена дослідженню способів вирішення проблеми дисбалансу класів у задачі класифікації страхових випадків. Для вирішення поставленого завдання було використано базу даних у сфері автострахування, в якій міститься інформація щодо наявності чи відсутності шахрайства за позовами клієнтів. Клас шахрайських випадків, який цікавить нас найбільше, представлений у базі втричі меншою кількістю записів за правомірні позови. Задля уникнення проблем моделювання на незбалансованих даних були застосовані методи передискретизації, зокрема випадковий оверсемплінг та SMOTE. Оцінка результатів, отриманих на різних вибірках, показує, що методи балансування дозволяють суттєво покращити якість класифікації.

У ході дослідження на отриманих наборах даних були побудовані класифікатори на основі логістичної регресії, методу опорних векторів, алгоритму k-найближчих сусідів, Байєсівського класифікатора, дерева рішень, випадкового лісу та нейронної мережі персептронного типу. Порівняльний аналіз показників якості побудованих класифікаторів допоміг визначити найкращі методи для виявлення шахрайських претензій. Для обох наборів даних такими методами були визнані логістична регресія та нейронна мережа, які мають високий рівень виявлення шахрайських випадків у поєднанні з належною загальною прогностичною силою моделі.

Ключові слова: *машинне навчання, нейронна мережа, логістична регресія, дерево рішень, класифікація, незбалансовані дані, оверсемплінг, випадковий оверсемплінг, SMOTE*

FRAUD DETECTION IN CAR INSURANCE: THE PROBLEM OF UNBALANCED SAMPLING

Kateryna Kononova

V.N. Karazin Kharkiv National University
4 Svobody Sq., Kharkiv, 61022, Ukraine
ORCID: 0000-0001-6990-5746, E-mail: kateryna.kononova@karazin.ua

Anna Havrylenko

V.N. Karazin Kharkiv National University
4 Svobody Sq., Kharkiv, 61022, Ukraine
ORCID: 0000-0002-9191-6238, E-mail: anna.gavrilenko99@gmail.com

Solving classification problems using machine learning methods, data scientists often face the problem of data imbalances. Class imbalance is common in financial sector, in particular for the task of fraud detection in car insurance. Training models on unbalanced data can lead to misclassifications and large numbers of false positives due to the tendency of the model to classify observed cases as the majority class.

This paper deals with the study of ways to solve the problem of class imbalance in the task of insurance claims classifying. To solve this problem, a database in the field of auto insurance was used, which provide information about the presence or absence of fraudulent customer claims. The class of fraudulent cases that interests us the most is represented in the database by three times fewer records than for legitimate claims. Oversampling techniques including random over-sampling and SMOTE were applied to avoid modeling problems on unbalanced data. Evaluation of the results obtained on different samples indicates that balancing methods can significantly improve the quality of the classification.

Logistic regression, support vector machine, k-nearest neighbors classifier, Bayesian classifier, decision tree, random forest and perceptron type neural network were built on the obtained datasets. A comparative analysis of the models' qualities allowed to determine the best methods for detecting fraudulent claims. For both datasets, logistic regression and neural network were recognized as such methods, having a high level of fraud detection combined with a good predictive power of the model.

Keywords: *machine learning, neural network, logistic regression, decision tree, classification, unbalanced data, oversampling, random over-sampling, SMOTE*

JEL Classification: C52, C55, G22

Постановка проблеми. Ринок страхування займає друге місце за рівнем капіталізації серед фінансових ринків України у небанківському секторі [7, 11].

Найпоширеніші види страхування на українському ринку – це автострахування (КАСКО, ОСЦПВ, «Зелена картка») та особисте страхування (життя та медичне). Проте і в Україні, і у світі саме автострахування є найнебезпечнішим для страховиків бізнесом – в деяких європейських країнах на КАСКО та ОСЦПВ припадає понад 80 % усіх махінацій.

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.

Найпоширенішим інструментом машинного навчання, що використовується в задачах виявлення шахрайства, є різноманітні класифікатори. Проте точність моделей класифікації напряму залежить від якості та складу даних, на яких вони побудовані. Через те, що частка негативних випадків та недоброчесних клієнтів значно менша, серйозною проблемою при вирішенні таких завдань стає значний дисбаланс класів. У разі небалансованої вибірки алгоритми машинного навчання, намагаючись ідентифікувати поодинокі випадки у досить великих наборах даних, мають тенденцію відносити їх до класу більшості, водночас даючи помилкове відчуття високоточної моделі [10].

Аналіз останніх досліджень. Для вирішення проблеми незбалансованої вибірки та покращення якості класифікації існують різні алгоритми, якими користуються дослідники, враховуючи вихідний набір даних та цілі побудови конкретного класифікатора. Методи та підходи до класифікації на незбалансованих наборах даних розглядають українські та

зарубіжні вчені, зокрема Білобородова та Скарга-Бандурова, які тестували мінімаксний підхід для лінійної бінарної класифікації даних [3], Савіна та Бень, які проблему незбалансованості даних вирішували шляхом пошуку ефективних процедур відбору пояснюючих змінних до моделі та формування ансамблів моделей [15], Каврін та Субботін, які запропонували метод обробки даних, що базується на поєднанні технологій андерсемплінгу (англ. undersampling – субдискретизація або видалення деякої кількості прикладів мажоритарного класу) та кластерного аналізу [9], Севастьянов та Щетинін, які для вирішення проблеми незбалансованих даних пропонували використовувати комбінацію алгоритмів класифікації та методів відбору ознак RFE (Recursive Feature Elimination – зворотне видалення ознак), а також проводили експерименти з алгоритмами випадкового лісу та Boruta (алгоритм автоматичного вибору ознак у наборі даних на основі випадкового лісу) з попереднім балансуванням класів методами випадкового семплювання, SMOTE (Synthetic Minority Oversampling TEchnique – спосіб передискретизації синтезованих меншин) та ADASYN (ADAptive SYNthetic sampling – адаптивне синтетичне формування вибірки) [16], Паклін, Уланов та Царьков, які показали, що з використанням оверсемплінгу (англ. oversampling – передискретизація або збільшення кількості прикладів міноритарного класу) можна побудувати ефективні моделі кредитного скорингу та обґрунтувати підбір порогового скорингового балу [12], Чаула, Суі, Ю, Гонг та Пан, які провели порівняльні дослідження різних методів семплінгу за їхньою ефективністю для покращення результатів класифікації методами машинного навчання [5, 17] та багато інших.

Особливості застосування методів передискретизації, зокрема вплив техніки оверсемплінгу SMOTE на результати побудови класифікаторів, вивчали Демидова та Клюєва [6], Патіл, Фреймвала та Казі [13], Фетласи, Охзахата, Ву та Като [14] та інші.

У страхуванні проблеми моделювання на незбалансованих вибірках з метою виявлення шахрайства розглянуто, наприклад, в роботі Хасана та Абрахама [8].

Як можна бачити з наведеного аналізу літературних джерел, проблематиці побудови ефективного класифікатора на незбалансованій вибірці приділяється значна увага у наукових дослідженнях та практичних розробках (беручи до уваги широкий перелік бібліотек, які реалізують алгоритми формування навчальних вибірок для побудови моделей із незбалансованих даних). Разом з тим, для різних прикладних задач та наборів даних ефективніше можуть проявляти себе різні математичні методи побудови моделей та алгоритми формування вибірок. Відповідно, проведемо дослідження з аналізу ефективності методів машинного навчання та алгоритмів семплінгу для вирішення практичної економічної задачі виявлення шахрайства в автострахуванні.

Мета та завдання. Мета дослідження полягає у вирішенні проблеми дисбалансу класів та порівнянні якості класифікаторів, побудованих на різних вибірках, на прикладі завдання виявлення шахрайства в автострахуванні.

Для дослідження страхових випадків методами машинного навчання була обрана база даних страхових претензій з автострахування, яка налічує 39 змінних та містить 1000 записів зі страхових претензій клієнтів. У датасеті [2] надана інформація щодо:

1) споживачів страхових послуг – 10 ознак, зокрема вік, рівень освіти, стать, вид зайнятості, хобі, сімейний стан та інші;

2) умов договору страхування – 7 ознак, зокрема ліміт поліса, франшиза, річна премія, межа покриття та інші;

3) характеристик інциденту – 22 ознаки, зокрема тип та важкість інциденту, тип зіткнення, факт звернення до органів влади, загальна сума претензії, марка та вік автомобіля, тілесні ушкодження та пошкодження майна, наявність поліцейського звіту та свідків.

Маючи на увазі мету роботи та структуру бази даних, нами були поставлені наступні завдання:

1) провести попередню підготовку даних,

2) побудувати базову модель класифікатора на вихідній вибірці,

3) провести низку експериментів щодо балансування даних,

4) побудувати класифікатори на основі алгоритмів машинного навчання,

5) виконати порівняльний аналіз результатів.

Всі завдання були реалізовані на високорівневій мові програмування Python, використовуючи інтерактивне середовище розробки Jupyter Notebook та спеціальні бібліотеки для аналізу даних.

Попередній аналіз та підготовка даних. Попередній аналіз даних показав, що з 39 змінних 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, а саме «property_damage», «police_report_available» та «collision_type», заповнено пропущені значення. Корегування викидів проведено у змінних «policy_annual_premium», «umbrella_limit» та «property_claim». Категоріальні змінні були закодовані. З метою забезпечення порівняного впливу факторів на ендогенну змінну, було здійснено шкалювання даних методом стандартизації.

Ендогенна змінна «fraud_reported», що визначає факт наявності чи відсутності шахрайства у страхових претензіях, представлена двома класами:

1) «Так» у кількості 247 випадків (встановлений факт шахрайства);

2) «Ні» у кількості 753 випадків, де шахрайство не виявлено.

Будуючи модель класифікатора, необхідно було визначити, яка подія є позитивною, а яка – негативною. Так як задача полягає у виявленні шахрайських випадків, то відповідно позитивною подією є наявність факту шахрайства, а негативною – його відсутність. Одразу слід відмітити, що співвідношення класів дорівнює приблизно 1:3, тобто ми маємо справу з незбалансованими даними.

Базова модель логістичної регресії. Для розв’язання завдання виявлення шахрайства в автострахуванні в якості базової моделі було обрано логістичну регресію. Для оцінки достовірності моделей використано одноразову перехресну перевірку, яка

реалізується розбиттям вибірки на дві взаємодоповнювані підвибірки: навчальну та тестову, у співвідношенні 70 на 30. Таким чином у навчальну вибірку потрапило 700 записів щодо страхових випадків, з яких лише 175 помічені як шахрайські, а в тестову – 300, 72 з яких є шахрайськими. Тобто клас шахрайських випадків, який цікавить нас найбільше, виступає класом меншості. У табл. 1 наведені результати тестування простого лінійного класифікатора, побудованого на незбалансованому наборі даних.

Таблиця 1

ПОКАЗНИКИ ЯКОСТІ БАЗОВОЇ МОДЕЛІ НА ТЕСТОВІЙ ВИБІРЦІ

Експеримент 1 (незбалансовані дані)		Модель		Показники якості моделі	
		Негативно	Позитивно	*Точність* (*Accuracy*)	89,00 %
Фактично	Ні	**212**	**16**	*Частка помилок* (*Error*)	11,00 %
				Чутливість (*Se*)	76,39 %
	Так	**17**	**55**	*Специфічність* (*Sp*)	92,98 %
				AUC	93,70 %

Джерело: авторська розробка

З табл. 1 бачимо, що передбачивши клас більшості, модель логістичної регресії демонструє доволі високі загальні оцінки якості. Тобто у цьому випадку ми спостерігаємо класичний «парадокс точності», коли висока точність моделі досягається за рахунок великої кількості правильно класифікованих негативних випадків (бо їх кількість у даних значно більша). Проте низька чутливість свідчить про те, що базова модель не надто якісно розпізнає випадки шахрайства, саме які цікавлять нас в першу чергу. На тестовій вибірці модель продукує практично рівну кількість хибнонегативних та хибнопозитивних прикладів. Але помилка першого роду (хибнонегативні приклади) становить основну загрозу для страховика через те, що подія, яка нас

цікавить (шахрайство), помилково не виявляється та класифікується як безпечна.

Балансування вибірки. Для подолання проблеми незбалансованих даних, навчання моделі необхідно виконувати з використанням вибірки з рівномірним розподілом класів. Це досягається шляхом застосування методів перебалансування. Суть методів передискретизації полягає або у видаленні елементів із занадто великого набору (андерсемплінг) та/або додаванні елементів в недостатньо великий набір (оверсемплінг) [1]. У нашій ситуації, коли датасет складається з досить невеликої кількості випадків, використовувати андерсемплінг не є доцільним, тож розглянемо результати моделювання з використанням оверсемплінгу.

В результаті застосування рандомного оверсемплінгу було отримано результати, що наведені у табл. 2.

Таблиця 2

ПОКАЗНИКИ ЯКОСТІ МОДЕЛІ НА ТЕСТОВІЙ ВИБІРЦІ

Експеримент 2 (випадковий оверсемплінг)		Модель		Показники якості моделі	
		Негативно	Позитивно	*Точність* (*Accuracy*)	87,00 %
Фактично	Ні	**196**	**32**	*Частка помилок* (*Error*)	13,00 %
				Чутливість (*Se*)	90,28 %
	Так	**7**	**65**	*Специфічність* (*Sp*)	85,96 %
				AUC	94,10 %

Джерело: авторська розробка

Бачимо, що в цьому випадку точність класифікатора трохи погіршала, що насправді добре, бо знижує хибні очікування від моделі. Проте помилка першого роду значно скоротилася, а чутливість моделі до розпізнавання випадків шахрайства виросла до 90 %, що є досить високим показником.

Проведемо ще один експеримент із застосуванням іншого методу оверсемплінгу. Цього разу скористаємося методом SMOTE, що на відміну від попереднього методу штучно створює нові елементи в безпосередній близькості від тих, що вже існують у меншому наборі. Аналізуючи по одному об'єкту за раз, SMOTE враховує різницю між спостереженням та його найближчим сусідом. Він зменшує різницю на випадкове число від нуля до одиниці та визначає нову точку, додаючи це значення до об'єкта. Таким чином, SMOTE не копіює спостереження, а натомість створює синтетичні дані на основі наявних [10].

Реалізуємо оверсемплінг, використовуючи бібліотеку imblearn (imbalanced-learn), що створена для подолання проблем, пов'язаних із незбалансованими наборами даних (рис. 1).

```
from imblearn.over_sampling import SMOTE
smote = SMOTE(sampling_strategy='minority',random_state=10)
X_sm, y_sm = smote.fit_resample(X, y)
```

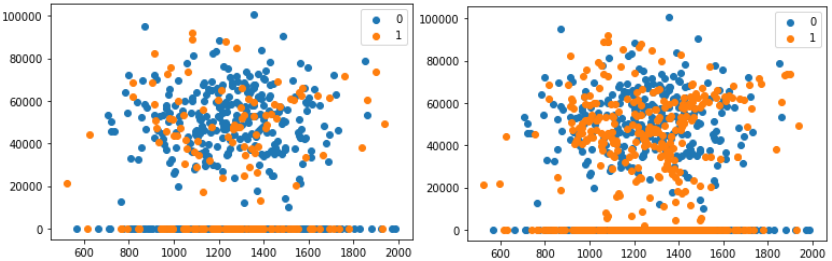

Рис. 1. Результат перебалансування методом SMOTE

Джерело: авторська розробка

З наведеної візуалізації можна побачити, що було створено 350 синтетичних шахрайських випадків, тож відтепер класи рівно збалансовані 1:1 (по 525 прикладів у кожному).

В результаті застосування методу SMOTE було отримано такі результати моделювання на тестовій вибірці (табл. 3).

Таблиця 3

ПОКАЗНИКИ ЯКОСТІ МОДЕЛІ НА ТЕСТОВІЙ ВИБІРЦІ

Експеримент 3 (SMOTE)		Модель		Показники якості моделі	
		Негативно	Позитивно	*Точність* (*Accuracy*)	86,66 %
Фактично	Ні	**195**	**33**	*Частка помилок* (*Error*)	13,33 %
				Чутливість (*Se*)	90,28 %
	Так	**7**	**65**	*Специфічність* (*Sp*)	85,52 %
				AUC	94,10 %

Джерело: авторська розробка

Бачимо, що в цьому випадку точність трохи погіршала, проте помилка першого роду та чутливість моделі не змінилися. Інтегральний показник якості моделі бінарної класифікації – площа під ROC кривою (показник *AUC* – Area Under Curve) в усіх трьох експериментах майже незмінна (рис. 2).

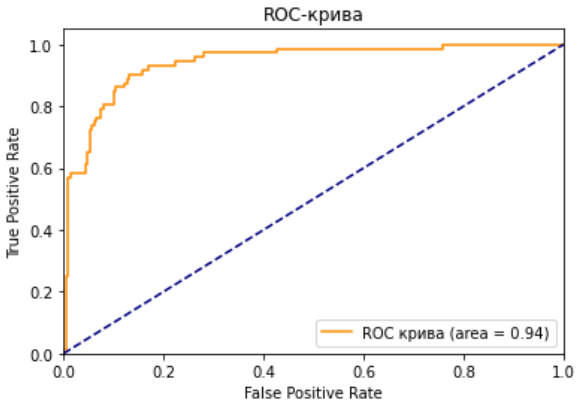

Рис. 2. ROC крива для моделі, побудованої із застосуванням методу SMOTE

Джерело: авторська розробка

Порівнюючи якість класифікаторів, побудованих на вихідній та перебалансованих вибірках, бачимо, що у разі застосування оверсемплінгу, моделі показують вищі значення чутливості, вони краще за базову модель виявляють шахрайські випадки. Хоча в експериментах 2 та 3 показник специфічності трохи знизився у порівнянні з базовою версію, його рівень залишається достатнім, а загальна прогностична сила цих моделей має високе значення (див. табл. 4). Отже, можна стверджувати, що реалізовані методи балансування виконують поставлену задачу та дають можливість покращити якість класифікації.

Таблиця 4

ПОРІВНЯННЯ ЯКОСТІ КЛАСИФІКАТОРІВ

Набір даних	Показник якості				
	Accuracy	***Error***	***Se***	***Sp***	***AUC***
Експеримент 1	0,890	0,110	0,764	0,930	0,937
Експеримент 2	0,870	0,130	0,903	0,859	0,941
Експеримент 3	0,866	0,130	0,902	0,855	0,941

Джерело: авторська розробка

Відбір значущих змінних. Продовжуючи експериментування з моделями, побудованими на збалансованих даних, проведемо відбір значущих змінних із застосуванням методів випадкового оверсемплінгу та SMOTE (див. табл. 5). Критерієм відбору значущих змінних був показник p-value[1], рівний 3 % (поріг було визначено в процесі експериментування). Змінні, за якими він перевищував це значення, не були використані для побудови моделей на значущих факторах. Для моделі, побудованої із

[1] p-value – це ймовірність істинності нульової гіпотези, що незалежні змінні не пояснюють динаміку залежної змінної. Якщо значення p-value нижче граничного рівня, то нульова гіпотеза помилкова. Один з найбільш поширених алгоритмів відбору інформативних ознак базується на розрахунку значення p-values, відповідних коефіцієнтам лінійної регресії.

застосуванням випадкового оверсемплінгу, було виявлено 13 значущих змінних, для SMOTE – 10. При цьому дев'ять змінних виявилися значущими для обох варіантів моделювання.

Таблиця 5

ПЕРЕЛІК ЗНАЧУЩИХ ЗМІННИХ

Випадковий оверсемплінг	**SMOTE**
insured_occupation – вид зайнятості insured_hobbies – хобі insured_relationship – сімейний стан capital–loss – втрата капіталу incident_type – тип інциденту collision_type – тип зіткнення incident_severity – важкість інциденту authorities_contacted – звернення до органів влади incident_state – район, у якому стався інцидент number_of_vehicles_involved – кількість задіяних транспортних засобів witnesses – свідки auto_make – марка автомобіля auto_model – модель автомобіля	policy_csl – єдиний комбінований ліміт поліса insured_occupation – вид зайнятості insured_hobbies – хобі insured_relationship – сімейний стан collision_type – тип зіткнення incident_severity – важкість інциденту incident_state – район, у якому стався інцидент number_of_vehicles_involved – кількість задіяних транспортних засобів witnesses – свідки auto_model – модель автомобіля

Джерело: авторська розробка

Класифікатори, побудовані з використанням інших алгоритмів. Проведемо порівняльний аналіз якості класифікаторів:

- LR – логістична регресія,
- SVM – метод опорних векторів,
- KNN – метод k-найближчих сусідів,
- NB – Байєсівський класифікатор,
- DT – дерево рішень,
- RF – випадковий ліс,
- NNP – нейронна мережа (двошаровий персептрон),

побудованих як на повному наборі змінних, так і на найбільш значущих факторах для кожного з методів – випадкового оверсемплінгу та SMOTE (табл. 6).

Таблиця 6

ПОРІВНЯЛЬНИЙ АНАЛІЗ ЯКОСТІ КЛАСИФІКАТОРІВ

Модель	Випадковий оверсемплінг					SMOTE				
	Accuracy	*Error*	*Se*	*Sp*	*AUC*	*Accuracy*	*Error*	*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
LR (значущі змінні)	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
SVM (значущі змінні)	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
KNN (значущі змінні)	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
NB (значущі змінні)	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
DT (значущі змінні)	0.870	0.145	0.944	0.842	0.913	0.870	0.145	0.944	0.842	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
RF (значущі змінні)	0.873	0.126	0.930	0.855	0.914	0.870	0.130	0.875	0.868	0.922
NNP (усі змінні)	0.863	0.136	0.931	0.840	0.940	0.876	0.123	0.902	0.868	0.941
NNP (значущі змінні)	0.890	0.110	0.916	0.880	0.940	0.880	0.120	0.931	0.864	0.945

Джерело: авторська розробка

Для даних, отриманих з використанням випадкового оверсемплінгу, за всіма алгоритмами кращі результати класифікації показують моделі, що були побудовані на значущих

факторах. Це вказує на правильність відбору показників, що впливають на виявлення шахрайства.

За сукупністю метрик найкращі результати продемонструвала двошарова нейронна мережа персептронного типу, побудована на 10 значущих змінних з лінійною функцією активації на прихованому шарі та сигмоїдальною на вихідному шарі.

Досить ефективно класифікують шахрайські випадки на цьому наборі даних також логістична регресія, метод опорних векторів, дерево рішень та метод випадкового лісу на значущих факторах. Найгірша якість класифікації спостерігається за методом Байєсівського класифікатора, а також k-найближчих сусідів, побудованих на усіх змінних. Метод KNN показує високу чутливість, у той час коли значення специфічності у порівнянні з іншими моделями нижче. Якщо ухвалювати рішення про використання певного методу, треба відштовхуватися не тільки від високого показника чутливості, а й від співвідношення різних оцінок. У рамках даної задачі нам важливіше виявити шахрайські випадки, але в той же час велике значення має відтворення реальної картини. Через це будемо орієнтуватись, в першу чергу, на високу частку визначення позитивних випадків у поєднанні з високим рівнем виявлення негативних випадків та загальну прогностичну силу моделі.

Аналіз результатів якості моделей, побудованих на даних, отриманих з використанням методу SMOTE, підтверджує необхідність відбору найбільш значущих змінних, в результаті чого якість класифікаторів стає вищою, ніж при включенні усіх показників до моделі.

Найкращими моделями, створеними на основі даних SMOTE, стали двошарова нейронна мережа типу персептрон, побудована на 13 значущих змінних з лінійною функцією активації на прихованому шарі та сигмоїдальною на вихідному шарі, та логістична регресія.

Високі показники якості також показала модель випадкового лісу, побудована на значущих змінних. Метод k-найближчих сусідів, що реалізовано на значущих змінних, добре визначає

випадки шахрайства, однак має нижчу специфічність та загальну точність, адже погано класифікує негативні випадки. Гірші результати показують Байєсівський класифікатор та метод k-найближчих сусідів на усіх змінних.

За результатами проведеного експериментального дослідження та аналізу якості моделей можна стверджувати, що найкращим рішенням при виборі варіанту класифікатора для виявлення шахрайських претензій є логістична регресія та нейронна мережа персептронного типу.

Якщо порівнювати результати класифікації за даними двох вибірок, то вища точність спостерігається при використанні методу SMOTE, де більшість шахрайських випадків є штучно створеними. Цей підхід є ефективним, оскільки нові приклади є близькими у просторі ознак до існуючих прикладів з класу меншості, проте недоліком цього підходу є те, що синтетичні приклади створюються без урахування класу більшості, що небезпечно у разі значного перетину класів [4].

Висновки. У роботі проведено дослідження ефективності методів передискретизації для вирішення проблеми незбалансованих даних. Через невеликій розмір вибірки, експериментування проводилося з використанням методів оверсемплінгу, зокрема випадкового оверсемплінгу та SMOTE. Обидва методи полягають у навчанні моделей на якомога більшій кількості прикладів міноритарного класу (шахрайських претензій в автострахуванні). Проте, на відміну від випадкового оверсемплінгу, при застосуванні якого вибірка доповнюється наявними даними, метод SMOTE полягає у синтетичному створенні випадків, наближених до шахрайських. Аналіз результатів моделювання показав, що методи балансування дозволяють значно покращити якість класифікації.

У ході дослідження на отриманих наборах даних були побудовані класифікатори на основі логістичної регресії, методу опорних векторів, алгоритму k-найближчих сусідів, Байєсівського класифікатора, дерева рішень, випадкового лісу та нейронної мережі. На основі порівняльного аналізу показників якості

класифікації було визначено найкращі алгоритми машинного навчання для виявлення шахрайських претензій. Для обох вибірок такими методами були визнані нейронна мережа персептронного типу та логістична регресія, які мають високу частку визначення випадків шахрайства у поєднанні із задовільною прогностичною якістю моделі.

Практичне значення отриманих результатів полягає в тому, що побудовані у ході дослідження моделі дають змогу з достатньою точністю класифікувати страхові випадки на доброчесні та шахрайські, тож їх можна рекомендувати до впровадження при автоматизації процесу виявлення шахрайських претензій. Аналіз значущих факторів може бути корисним при ухваленні рішень щодо підписання угод з певними клієнтами та виплати їм страхового відшкодування.

References

1. Agarwal, R. (2019, July 21). *The 5 Sampling Algorithms every Data Scientist need to know*. Towards Data Science. https://towardsdatascience.com/the-5-sampling-algorithms-every-data-scientist-need-to-know-43c7bc11d17c

2. Alencar, R. (2017). *Resampling strategies for imbalanced datasets* [Data set]. Kaggle. Retrieved April 5, 2020, from https://www.kaggle.com/rafjaa/resampling-strategies-for-imbalanced-datasets#t3

3. Biloborodova, T., & Skarga-Bandurova, I. (2017). Pidkhody do klasyfikatsii nezbalansovanykh i zsunutykh naboriv danykh [Approaches for Classification of Imbalanced and Skewed Datasets]. *Visnyk Skhidnoukrainskoho natsionalnoho universytetu imeni Volodymyra Dalia (Bulletin of Volodymyr Dahl East Ukrainian National University), 8*(238), 17-24. [in Ukrainian]

4. Brownlee, J. (2020, January 17). *SMOTE for Imbalanced Classification with Python*. Machine Learning Mastery. https://machinelearningmastery.com/smote-oversampling-for-imbalanced-classification/

5. Chawla, N.V. (2009). Data mining for imbalanced datasets: An overview. In O. Maimon, & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (pp. 875-886). Springer. https://doi.org/10.1007/978-0-387-09823-4_45

6. Demidova, L. A., & Klyueva, I. A. (2017). Alhoritm podbora znacheniy parametrov bSMOTE-algoritma v zadache SVM-klassifikatsii na osnove nesbalansirovannykh naborov dannykh [Search algorithm of the parameters values of the bSMOTE-algorithm in the problem of the SVM classification based on the imbalanced datasets]. *Vestnyk Riazanskoho hosudarstvennoho radyotekhnycheskoho unyversyteta (Vestnik of Ryazan State Radio Engineering University), 61*, 67-77. http://vestnik.rsreu.ru/images/archive/2017/3-61/3.1__.pdf [in Russian]

7. FORINSURER. (2020, May 7). *Statystyka strakhovoho rynku Ukrainy [Statistics of the insurance market of Ukraine]*. https://forinsurer.com/stat [in Ukrainian]

8. Hassan, A.K.I., & Abraham, A. (2016). Modeling Insurance Fraud Detection Using Imbalanced Data Classification. In N. Pillay, A. Engelbrecht, A. Abraham, M. du Plessis, V. Snášel, & A. Muda (Eds.), *Advances in Intelligent Systems and Computing: Vol. 419. Advances in Nature and Biologically Inspired Computing* (pp. 117-127). Springer. https://doi.org/10.1007/978-3-319-27400-3_11

9. Kavrin, D. A., & Subbotin, S. A. (2018). Metody kolichestvennogo resheniya problemy nesbalansyrovannosti klassov [The methods for quantitative solving the class imbalance problem]. *Radioelektronika, informatyka, upravlinnia (Radio Electronics, Computer Science, Control), 1*, 83-90. https://doi.org/10.15588/1607-3274-2018-1-10 [in Russian]

10. Lahera, G. (2019, January 22). *Unbalanced Datasets & What To Do About Them*. Strands Tech Corner. https://medium.com/strands-tech-corner/unbalanced-datasets-what-to-do-144e0552d9cd

11. National Bank of Ukraine. (2020). *Register of Reporting Indicators for Non-Bank Financial Institutions* [Data set]. Retrieved May 7, 2020, from https://bank.gov.ua/ua/statistic/nbureport/statreport-nonbanking

12. Paklin, N. B., Ulanov, S. V., & Tsarkov, S. V. (2010). Postroyeniye klassifikatorov na nesbalansirovannykh vyborkakh na primere kreditnogo skoringa [Classifiers Construction Based on Imbalanced Datasets by the Example of Credit Scoring]. *Shtuchnyi Intelekt (Artificial Intelligence), 49*(3), 528-534. https://jai.in.ua/index.php/%D0%B0%D1%80%D1%85%D1%96%D0%B2?paper_num=984 [in Russian]

13. Patil, A., Framewala, A., & Kazi, F. (2020). Explainability of SMOTE Based Oversampling for Imbalanced Dataset Problems. In *Proceedings of 2020 3rd International Conference on Information and Computer Technologies* (pp. 41-45). IEEE. https://doi.org/10.1109/ICICT50521.2020.00015

14. Phetlasy, S., Ohzahata, S., Wu, C., & Kato, T. (2019). Applying SMOTE for a sequential classifiers combination method to improve the performance of intrusion detection system. In *Proceedings of 2019 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress* (pp. 255- 258). IEEE. https://doi.org/10.1109/DASC/PiCom/CBDCom/CyberSciTech.2019.00054

15. Savina, S., & Ben', V. (2015). Obiednannia modelei logit-rehresiy yak komitetu ekspertiv dlia otsinky kredytospromozhnosti pozychalnyka [Integration of models of logit-regressions as a committee of experts to assess the creditworthiness of borrower]. *Neiro-Nechitki Tekhnolohii Modelyuvannya v Ekonomitsi (Neuro-Fuzzy Modeling Techniques in Economics)*, *4*, 154-188. https://doi.org/10.33111/nfmte.2015.154 [in Ukrainian]

16. Sevastianov, L. A., & Shchetinin, E. Yu. (2020). O metodakh povysheniya tochnosti mnogoklassovoy klassifikatsiyi na nesbalansirovannykh dannykh [On methods for improving the accuracy of multiclass classification on imbalanced data]. *Informatika i ieyo primeneniya (Informatics and its applications)*, *1*(14), 63-70. https://doi.org/10.14357/19922264200109 [in Russian]

17. Sui, Y., Yu, M., Hong, H., & Pan, X. (2019). Learning from imbalanced data: A comparative study. In W. Meng, & S. Furnell (Eds.), *Communications in Computer and Information Science: Vol. 1095. Security and Privacy in Social Networks and Big Data* (pp. 264-274). Springer. https://doi.org/10.1007/978-981-15-0758-8_20

Стаття надійшла до редакції 06.08.2020