УДК 004:519.2
Т. Базан
(Тернопільський національний технічний університет імені Івана Пулюя, Україна)

АНАЛІЗ ВХІДНИХ ДАНИХ СИСТЕМИ ПРОГНОЗУВАННЯ ФІНАНСОВОЇ РЕНТАБЕЛЬНОСТІ ПІДПРИЄМСТВА

UDC 004:519.2
T. Bazan

ANALYSIS OF INPUT DATA OF THE SYSTEM FOR FORECASTING THE FINANCIAL PROFITABILITY OF THE ENTERPRISE

Для виконання дослідження використовуються дані із сайту Kaggle [1] – відкритого джерела різноманітних конкурсних завдань у сфері аналізу даних. Дана платформа дозволяє ділитися своїми рішеннями, обговорювати процес аналізу та змагатися у точності побудованих моделей. Використані дані «Financial Distress Prediction: Bankruptcy Prediction». Контекст даних – це набір даних (НД), призначений для прогнозування фінансової кризи для вибірки компаній. Фрагмент стовпців даних представлений на рисунку.

	Company	Time	Financial Distress	x1	x2	x3	x4
0	1	1	0.010636	1.2810	0.022934	0.87454	1.21640
1	1	2	-0.455970	1.2700	0.006454	0.82067	1.00490
2	1	3	-0.325390	1.0529	-0.059379	0.92242	0.72926
3	1	4	-0.566570	1.1131	-0.015229	0.85888	0.80974
4	2	1	1.357300	1.0623	0.107020	0.81460	0.83593
...	...	...	...	...	...	...	...

Перший стовпець: Є ідентифікатором компанії. Другий стовпець: Показує різні періоди часу, до яких належать дані. Довжина часового ряду варіюється від 1 до 14 кожної компанії. Третій стовпець: Цільова змінна позначається як «Фінансова криза», якщо вона більша за -0,5, то компанію слід вважати рентабельною (0). В іншому випадку він був би розцінений як фінансово неблагополучний (1). Від четвертого до останнього стовпця: характеристики, позначені від x1 до x83, є деякі фінансові та нефінансові характеристики відібраних компаній. Ці характеристики відносяться до попереднього періоду часу, який слід використовувати для прогнозування того, буде компанія відчувати фінансові труднощі чи ні. Класифікація. Особливість x80 є категоріальною змінною. Наприклад, компанія 1 зазнає фінансових труднощів у момент 4, але компанія 2 все ще знаходиться в хорошому стані в момент 14.

Цей НД не є збалансованим (є 136 компаній із фінансовими проблемами проти 286 здорових, тобто 136 спостережень за звітний рік є фінансовими проблемами, тоді як 3546 спостережень за звітний рік є рентабельними). Слід зазначити, що 30% цього набору даних слід випадковим чином призначити як набір тестових даних, так що 70% тих, що залишилися, використовуються для вибору функцій і вибору моделі.

До даних також додаються примітки. Стовпець перший – ці дані можна розглядати як проблему класифікації. Стовпець другий – дані також можна розглядати як проблему регресії, а потім результат буде перетворено на класифікацію. Стовпець третій – ці дані можна розглядати як багатовимірну класифікацію часових рядів.

Основними проблемами даних для дослідження є такі пункти: які особливості найбільше свідчать про фінансову скруту; які типи моделей машинного навчання найкраще працюють із цим набором даних?

Література

1. Kaggle: Your Machine Learning and Data Science Community. URL: ^https://www.kaggle.com/ (дата звертання: 01.12.22).