

ВИКОРИСТАННЯ МЕТОДІВ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ПІДВИЩЕННЯ ТОЧНОСТІ КЛАСИФІКАЦІЇ СПОЖИВАЧІВ КРЕДИТІВ

Здатність передбачати майбутнє завжди цінувалася дорого. Зараз прогнозування поставлено на твердий математичний ґрунт. І банки користуються цим, впроваджуючи системи автоматизованого скорингу, призначені для оцінки кредитних ризиків фізичних осіб.

Скорингові системи – популярний в банківській сфері інструмент оцінки клієнтів, в основі якого закладені статистичні методи. Такі системи дозволяють знизити витрати і мінімізувати операційний ризик за рахунок автоматизації прийняття рішення про надання кредиту [1].

Найбільш популярними сьогодні є методи побудови скорингових алгоритмів на основі логістичної регресії, дерева класифікації та нейронної мережі [2].

У ході виконання роботи вирішувалося питання розв'язання задачі кредитного скорингу за допомогою алгоритму машинного навчання, заснованого на дереві пошуку рішень, який використовує фреймворк градієнтного бустинга – XGBoost. Він забезпечує високоєфективну реалізацію алгоритму посиленням стохастичного градієнта та доступом до набору параметрів моделі, призначених для забезпечення контролю над процесом навчання моделі [3].

Для адекватної роботи моделі машинного навчання необхідні чітко визначені вихідні дані і інструкції. Вихідні дані часто можуть включати в себе шуми, викиди чи відсутні значення. Саме тому у Data Mining виділяють підготовку даних (Data Preparation, Preprocessing) у окрему фазу, що може займати до 80% всіх витрат ресурсів і часу. Препроцесінг даних включає наступні задачі опрацювання вихідних даних:

- вибірка даних – відбір ознак (features або предикторів) і об'єктів з урахуванням їх релевантності для цілей Data Mining;
- очищення даних – видалення помилок, неправильних та відсутніх значень, дублікатів, відновлення цілісності і логічних зв'язків;
- генерація ознак – відбір ознак, їх трансформація для підвищення точності алгоритмів машинного навчання.

Після проведення всіх вищезгаданих операцій, була побудована модель алгоритму XGBoost. Наслідком пошуку найбільш релевантних для вихідних даних параметрів стала побудова нової моделі, точність прогнозу якої на порядок перевищує отримані раніше результати. Ця модель і була застосована для розв'язання задачі кредитного скорингу.

¹ Студентка групи 124м-19, НТУ «Дніпровська політехніка»

² д.ф.-м.н., проф. кафедри САУ, НТУ «Дніпровська політехніка»

	id	pred	
	144	5388	1
	270	5514	1
	563	5807	1
	607	5851	1
	684	5928	1

	9048	6435	1
	9436	10675	1
	9593	8559	1
	9722	8663	1
	9850	6537	1

107 rows × 2 columns

Рисунок 1 – Результат розв’язку задачі (позитивні передбачення)

З рисунку 1 видно, що розв’язок полягає у 107 схваленому кредиті з 9922 можливих, тобто лише трохи більше одного відсотку (~1,078) позичальників отримало схвалення (під час тесту співвідношення було 144 до 7880, тобто ~1,827%).

Висновки

Алгоритм XGBoost показав гарні результати для вирішення задачі кредитного скорингу, та при більш ретельному виборі параметрів результати прогнозу можуть дати кращий результат і підвищити кількість виданих кредитів. В такому випадку треба орієнтуватися на пріоритети банку – цілком можливо, що помилкове схвалення кредиту може нанести більший збиток, ніж недоотриманий прибуток від помилкової відмови.

ПЕРЕЛІК ПОСИЛАНЬ

1. Малюгін В., Грінь Н. // Про ефективність статистичних алгоритмів кредитного скорингу // Банкаўські веснік, 2010, 39 с.
2. Jonathan Hirko // Intro to Classification and Feature Selection with XGBoost // AI Time Journal. 2019.
3. Дьяконов А.Г. // Методи вирішення задач класифікації з категоріальними ознаками // Прикладна математика та інформатика, 2014, 11 с.