УДК 631.559.8

А.О. Лавренюк

ДОСЛІДЖЕННЯ МОДЕЛЕЙ ІНТЕЛЕКТУАЛЬНОГО АНАЛІЗУ ДАНИХ ДЛЯ ПОБУДОВИ СИСТЕМ КРЕДИТНОГО СКОРИНГУ

Вінницький національний технічний університет

Анотація

Проведено порівняння різних методів інтелектуального аналізу даних (ІАД) при побудові моделей кредитного скорингу. Визначено основні переваги і недоліки методів ІАД на різних етапах розробки систем кредитного скорингу. Попереднє тестування показало, що найкращі результати демонструють логістична регресія і випадковий ліс, а метод головних компонент виявився найпростішим і найшвидшим засобом аналізу даних.

Ключові слова: *кредитний скоринг, інтелектуальний аналіз даних, логістична регресія, випадковий ліс, метод опорних векторів*

Abstract

In the paper, a comparison of different data mining methods for building credit scoring models has been conducted. Data mining methods' main advantages and disadvantages at various stages of credit scoring system development have been identified. Preliminary testing showed that logistic regression and random forest demonstrate the best results, while the principal component method proved to be the simplest and fastest means of data analysis.

Keywords: credit scoring, data mining, logistic regression, random forest, support vector machine

Вступ

Кредитним скорингом називається методологія оцінювання кредитоспроможності потенційних позичальників, класифікації потенційних клієнтів банку по рівню ризику, набір моделей прийняття рішень щодо вирішення питання надання споживчого кредиту. Найчастіше постановкою задачі для скорингу в поняттях ризику контрагента є прогнозування показників, які є ранніми індикаторами по відношенню до дефолту, але призводять до нього з високою ймовірністю.

Швидкий розвиток машинного навчання і заснованого на ньому інтелектуального аналізу даних (ІАД) надав нові методологічні можливості для побудови вдосконалених моделей кредитного скорингу, які можуть базуватись щонайменше на шести його класичних задачах: класифікації, кластеризації, прогнозуванні, оцінюванні, асоціації та візуалізації, – і використовуються на етапах кореляційного аналізу і аналізу прогностичної сили вхідних змінних, при постановці задачі моделювання, побудові моделі та оцінюванні якості прогнозів. Дуже важливими для побудови моделей кредитного скорингу є і методи попередньої обробки даних ІАД. Отже, значний інтерес викликає дослідження ефективності застосування як окремих методів ІАД, так і їх комбінації на різних етапах кредитного скорингу. Необхідно здійснити порівняльний аналіз методів ІАД і визначити основні переваги і недоліки методів ІАД при їх використанні на різних етапах побудови систем кредитного скорингу.

Результати дослідження

Для порівняння ефективності різних методів інтелектуального аналізу даних (ІАД) в задачі побудови моделей кредитного скорингу були підготовлені набори даних з архіву центру машинного навчання та інтелектуальних систем [1]. Ці набори даних містили інформацію про позичальників, зокрема їх соціально-демографічні характеристики, кредитну історію та інші релевантні фактори. З метою зменшення розмірності даних та відбору найбільш значущих параметрів були застосовані метод головних компонент і випадковий ліс. Метод головних компонент дозволив виділити приховані закономірності в даних і сформувати нові змінні, які пояснюють більшу частину варіації в початкових даних. Випадковий ліс, в свою чергу, дав можливість оцінити важливість кожної змінної для прогнозування цільової змінної (дефолту позичальника) і відібрати найбільш інформативні предиктори. В результаті застосування цих методів були отримані два зменшені набори даних, які використовувались для подальшого аналізу і побудови моделей.

Для порівняння ефективності різних підходів до побудови моделей кредитного скорингу були відібрані наступні методи: метод опорних векторів, логістична регресія, бустинг і випадковий ліс. Ці методи представляють різні сім'ї алгоритмів машинного навчання і мають свої особливості в контексті вирішення задачі класифікації. Метод опорних векторів базується на ідеї розділення класів в багатовимірному просторі ознак за допомогою гіперплощини, яка максимізує відстань між найближчими представниками різних класів. Логістична регресія є імовірнісним методом, який моделює залежність ймовірності належності об'єкта до певного класу від значень його ознак. Бустинг і випадковий ліс відносяться до ансамблевих методів, які комбінують множину базових моделей для отримання більш точного і стабільного прогнозу. Крім того, отримані результати планується порівняти з результатами попередніх досліджень, в яких використовувались генетичні алгоритми і нечіткі нейронні мережі [2, 3]. Це дозволить оцінити ефективність обраних методів в порівнянні з іншими підходами, які довели свою ефективність в задачах кредитного скорингу.

Одним з ключових етапів порівняння моделей машинного навчання є вибір гіперпараметрів кожної моделі, які не можуть бути оптимізовані на навчальному наборі даних і мають бути визначені апріорно. Гіперпараметри, такі як коефіцієнт регуляризації, кількість дерев в ансамблі, глибина дерев тощо, можуть суттєво впливати на якість моделі і мають бути ретельно підібрані для конкретної задачі. Для вирішення цієї проблеми було застосовано метод крос-валідації, який дозволяє оцінити якість моделі на незалежному наборі даних і вибрати оптимальні значення гіперпараметрів. Крос-валідація передбачає розбиття набору даних на k частин, послідовне навчання моделі на k-1 частинах і тестування на частині, що залишилась. Цей процес повторюється k разів, і результати усереднюються для отримання більш стабільної оцінки якості моделі.

Для оцінювання точності визначення як надійних, так і потенційно проблемних позичальників використовувався графік ROC (Receiver Operating Characteristic). Цей інструмент дозволяє оцінити співвідношення між часткою правильно класифікованих позитивних прикладів (надійних позичальників) і часткою неправильно класифікованих негативних прикладів (проблемних позичальників) при різних порогових значеннях ймовірності дефолту. Площа під ROC-кривою (AUC) є агрегованою метрикою якості моделі, яка дозволяє порівнювати різні моделі між собою. Чим більше значення AUC, тим кращою є дискримінаційна здатність моделі.

За результатами попереднього тестування моделей на різних наборах даних найкращу ефективність продемонстрували логістична регресія і випадковий ліс. Ці моделі показали високі значення AUC (понад 0.8) і збалансовану точність по обох класах позичальників. При цьому випадковий ліс показав дещо кращі результати в порівнянні з логістичною регресією, особ ливо при тестуванні на наборах даних з різних джерел. Це може бути пов'язано з більшою гнучкістю і здатністю випадкового лісу моделювати складні нелінійні залежності в даних. Однак, слід зазначити, що випадковий ліс потребує більше обчислювальних ресурсів і часу для навчання і застосування моделі в порівнянні з логістичною регресією. В умовах обмежених ресурсів або необхідності швидкого прийняття рішень логістична регресія може бути більш прийнятним вибором.

Висновки

Отримані результати свідчать про перспективність використання методів ІАД, зокрема логістичної регресії і випадкового лісу, для побудови ефективних моделей кредитного скорингу. Ці методи дозволяють з високою точністю прогнозувати ймовірність дефолту позичальника на основі наявної інформації і можуть бути використані для оптимізації процесу прийняття кредитних рішень в банках та інших фінансових установах. Подальші дослідження можуть бути спрямовані на вдосконалення методології порівняння моделей, зокрема в частині вибору оптимальних наборів вхідних змінних, застосування більш складних методів попередньої обробки даних і ансамблевих підходів до побудови моделей.

Крім того, важливим напрямком подальших досліджень є врахування специфіки конкретних банків і сегментів позичальників при побудові моделей кредитного скорингу. Різні фінансові установи можуть мати різні вимоги до точності моделей, швидкості прийняття рішень, інтерпретабельності результатів тощо. Тому вибір оптимального методу або комбінації методів має здійснюватись з урахуванням цих факторів. Також, моделі кредитного скорингу мають періодично переглядатись і оновлюватись з урахуванням змін в економічній ситуації, поведінці позичальників і доступних даних.

Ще одним перспективним напрямком є інтеграція моделей кредитного скорингу з іншими системами підтримки прийняття рішень в банках, такими як системи виявлення шахрайства, оцінки

ризиків, маркетингу тощо. Це дозволить отримати більш повну картину кожного позичальника і прийняти більш зважене рішення щодо надання кредиту.

СПИСОК ВИКОРИСТАНОЇ ЛІТЕРАТУРИ

1. Machine Learning Repository UCI [Електронний ресурс]. – Режим доступу до ресурсу: http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
2. Brown, I., & Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. Expert Systems with Applications [Електронний ресурс]. – Режим доступу до ресурсу: https://doi.org/10.1016/j.eswa.2011.09.033
3. Kaminsky A.B., Pysanets K.S. Credit Scoring Model Based on Neural Networks and Genetic Algorithms // Proceedings of the 2019 IEEE International Conference on Advanced Trends in Information Theory (ATIT). – Kyiv, Ukraine, 2019. – P. 274-279.

Лавренюк Андрій Олегович — студент групи ЗАКІТР-23м, Факультет комп'ютерних систем і автоматики, Вінницький національний технічний університет, Вінниця, e-mail: andretti_alo@hotmail.com
Науковий керівник - ***Ковтун В'ячеслав Васильович*** — д-р техн. наук, професор, завідувач кафедри Комп`ютерних систем управління, Вінницький національний технічний університет, м. Вінниця

Andrii Lavreniuk O. — student of group ЗАКІТР-23m, Faculty of Computer Systems and Automation, Vinnytsia National Technical University, Vinnytsia, e-mail: andretti_alo@hotmail.com
Supervisor - ***Kovtun Vyacheslav V.*** — Dr. Tech. Sciences, professor, head of the Department of Computer Management Systems, Vinnytsia National Technical University, Vinnytsia