УДК 681.518.25

# РЕАЛІЗАЦІЯ АЛГОРИТМУ РОЗРАХУНКУ ПОКАЗНИКА ДЖИНІ ТА СТАТИСТИКИ КОЛМОГОРОВА-СМИРНОВА ДЛЯ БІНАРНОГО КЛАСИФІКАТОРА ЗАСОБАМИ МОВИ SQL

*Солошенко Олександр, Бідюк Петро*

Національний Технічний Університет України «Київський Політехнічний Інститут»

## Анотація

*Метою дослідження є реалізація розрахунку показника Джині та статистики Колмогорова-Смирнова для довільного бінарного класифікатора засобами мови SQL. Методикою реалізації є застосування табличних виразів, агрегатних та аналітичних функцій, з'єднань таблиць, множинних операцій та інших засобів DML мови SQL на прикладі Oracle Database 11g.*

## Abstract

*The objective of the research is the implementation of the estimation the GINI indicator and the Kolmogorov-Smirnov statistics for an abstract binary classifier using SQL language possibilities. The implementation methodologies are the application of the common table expressions, aggregate and analytic functions, table joins, set operations and other possibilities of the DML of the SQL using Oracle Database 11g as an example.*

## Вступ

У задачах бінарної класифікації даних, зокрема у задачах кредитного скорингу [1], необхідно оцінювати якість бінарної класифікації при змінному порозі відсікання, що накладається на присвоєний ймовірнісний результат приналежності до одного класу або на однозначно відповідний ймовірності присвоєний скоринговий бал. Індикатор Джині (GINI) та статистика Колмогорова-Смирнова (K.-S.) оцінюють якість бінарного класифікатора [2]. Актуальність полягає у необхідності розробки коду підрахунку показників якості бінарного класифікатора для систем керування базами даних (СКБД), що підтримують аналітичні функції з конструкцією «over». Головним чинником актуальності дослідження є збереження даних як таблиць СКБД в більшості організацій.

## Постановка задачі

Реалізація коду підрахунку індикатора Джині та статистики Колмогорова-Смирнова засобами мови маніпулювання даними (Data Manipulation Language, DML) мови структурованих запитів (Structured Query Language, SQL), тобто мовою програмування четвертого покоління (Fourth-Generation programming Language, 4GL).

## Способи обчислення GINI та K.-S.

Введемо позначення одного класу як B (bad) – негативного, іншого як G (good) – позитивного, виходячи з порівняння математичного очікування по рангу.

Статистика Колмогорова-Смирнова (K.-S.) обчислюється як максимальна абсолютна різниця значень функцій розподілу класів на області визначення рангу [1]:

$$KS = \max_{x \in X} | F_B(x) - F_G(x) | .$$

Щодо показника Джині (GINI), то індикатор обчислюється як інтеграл з вирахуванням інтегралу від «лінії байдужості» співвіднесений до максимально можливого значення площі над діагоналлю [3]. Існує три способи підрахунку GINI:

1) відношення площі фігури над діагоналлю до прямокутного трикутника площею 0,5 на параметричному графіку від порогу відсікання для

функції розподілу позитивних спостережень по осі абсцис, негативних – по осі ординат;

2) операційна характеристика розпізнавача (Receiver Operating Characteristics) [1];

3) крива Лоренца функції розподілу негативних спостережень від долі вибірки.

Згідно з пунктом 1, формулу індексу Джині можна записати таким чином:

$$GINI = \left( \int_{x \in X} F_B(x) dF_G(x) - \frac{1}{2} \right) / \left( \frac{1}{2} \right),$$

де інтеграл обчислюється методом трапецій по унікальних значеннях рангу:

$$\int_{x \in X} F_B(x) dF_G(x) = \sum_{x_i \in X} \frac{(F_B(x_i) + F_B(x_{i-1}))}{2} (F_G(x_i) - F_G(x_{i-1})),$$

де множина значень функцій розподілу доповнюється нулями для деякого «нульового» параметру, щоб перші значення функцій розподілу мали попереднє нульове значення (рис. 1). Дану криву теж можна назвати кривою Лоренца (але параметричною).

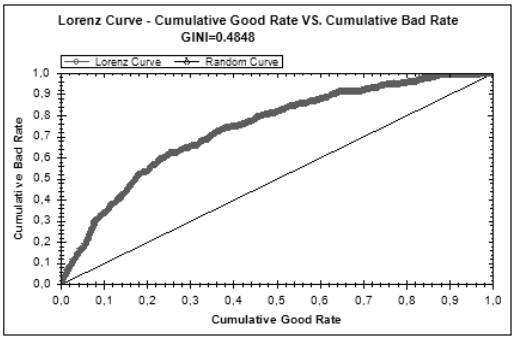


Рисунок 1 – Функції розподілу двох класів як параметричний двовимірний графік

## Реалізація обчислення GINI та K.-S. мовою SQL

Мова структурованих запитів (SQL) відноситься до мов програмування четвертого рівня (4GL) [4]. Потужним засобом більшості реляційних СКБД є віконні агрегатні та аналітичні функції [5]. У даній реалізації використовується функція LAG OVER().

Код реалізації (BAD та GOOD набувають значень 0 або 1, BAD + GOOD = 1):

```
WITH smpl(BAD, GOOD, score) AS (
  /*start sample*/
  SELECT BAD, 1 - BAD AS GOOD, SCOR_VAL AS SCORE
  FROM    input_table
  /*end sample*/
),
distr AS (
  SELECT score, (sum(GOOD)/sum(sum(GOOD)) over()) AS GOOD,
                (sum(BAD) /sum(sum(BAD))  over())    AS BAD
  FROM smpl
  GROUP BY score
  ORDER BY score
),
cum AS (
```

```
 SELECT   D_BASE.SCORE, sum(D_LESS.GOOD) AS GOOD,
                                sum(D_LESS.BAD)   AS BAD
 FROM distr d_base
         LEFT OUTER JOIN
         distr d_less ON D_LESS.SCORE<=D_BASE.SCORE
 GROUP BY D_BASE.SCORE
 ORDER BY D_BASE.SCORE
),
cum_with_lag AS (
 SELECT cum.*, LAG(cum.GOOD, 1, 0) OVER(ORDER BY SCORE) AS GOOD_PREV,
                    LAG(cum.BAD, 1, 0)   OVER(ORDER BY SCORE) AS BAD_PREV
 FROM cum
)
SELECT 'K.-S.' AS "indicator",
          ROUND(100.0*100.0*max(abs(BAD-GOOD)))/100.0||'%' AS "value"
FROM cum
UNION ALL
SELECT  'GINI' AS "indicator",
ROUND(100.0*100.0*
((sum((GOOD-GOOD_PREV)*(BAD+BAD_PREV)/2)-0.5)/0.5))/100.0||'%' AS "value"
FROM cum_with_lag;
```

Запропонована реалізація є лаконічною (завдяки віконній аналітичній функції), узагальненою завдяки табличним виразам, де користувачу необхідно замінити вміст табличного виразу позначеного «smpl». Приклад формату виведення наведено в табл. 1.

Таблиця 1 – Приклад формату виведення при виконанні коду

| Indicator | Value |
|---|---|
| K.-S. | 37,22% |
| GINI | 48,48% |

## Висновки

Запропонована реалізація застосовна для перевірки якості прогнозів бінарного класифікатора. Ключові показники, реалізовані в роботі, мають застосування в системах оцінювання якості роботи скорингових моделей у фінансовому моделюванні [6].

Перспективи подальших досліджень включають перенесення даного готового коду в пакетну процедуру PL/SQL або процедуру T-SQL конкретних реляційних СКБД.

## Список використаних джерел:


1. Lyn C. Thomas, David B. Edelman, Jonathan N. Crook. Credit Scoring and its Applications: SIAM monographs on mathematical modeling and computation. – University City Science Center, Philadelphia, SIAM, 2002. – 248 p. – ISBN 0-89871-483-4.
2. Naeem Siddiqi. Credit risk scorecards: developing and implementing intelligent credit scoring. – Hoboken: John Wiley & Sons, Inc., 2006. – 196 p. – ISBN 978-0-471-75451-0.
3. Руководство по кредитному скорингу: Учебн. пособие / Под ред. Элизабет Мэйз; пер. с англ. И.М. Тикота; науч. ред. Д.И. Вороненко. – Минск: Гревцов Паблишер, 2008. – 464 с. – ISBN 978-985-6569-34-3.
4. Урман Скотт. ORACLE 8. Программирование на языке PL/SQL. – Москва: Лори, 1999. – 607 с. – ISBN 5-85582-043-2.

5. Ицик Бен-Ган. Microsoft SQL Server 2012. Высокопроизводительный код T-SQL. Оконные функции. – Санкт-Петербург: БХВ-Петербург, 2013. – 256 с. – ISBN 978-5-7502-0416-8.

6. Бідюк П.І., Кузнєцова Н.В., Терентьєв О.М. Система підтримки прийняття рішень для аналізу фінансових даних // Наукові вісті НТУУ "КПІ". – 2011. – № 1. – С. 48–61. – ISSN 1810-0546.