Статистичні методи оцінки кредитоспроможності позичальника комерційного банку

Виконала:
студентка гр. КА-21 ІПСА
Пилєва Ганна Валеріївна

Науковий керівник:
д.т.н., проф. Бідюк Петро Іванович

Об’єкт, предмет , мета і метод дослідження

- ***Об`єкт дослідження:*** кредитний ризик банку, представлений історичними даними про його позичальників за певний операційний період.

- ***Предмет дослідження:*** кредитний скоринг як статистичний метод оцінки кредитоспроможності позичальника комерційного банку.

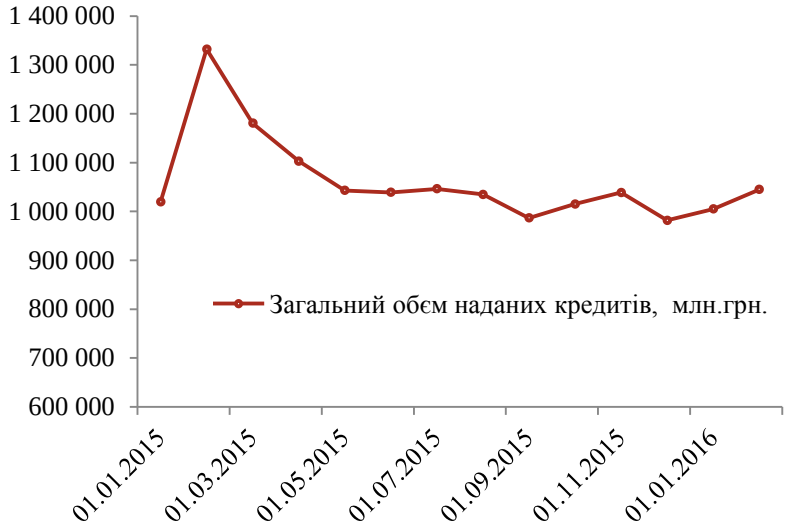
- ***Мета роботи:*** створення скорингової системи для оцінки кредитоспроможності нових позичальників за наявними історичними даними.

- ***Метод дослідження*** : аналіз методів класифікації на предмет можливості та ефективності їх застосування у процедурі кредитного скорингу.

Актуальність теми

Попередження або максимальне зменшення кредитного ризику залишається актуальним, а методологія оцінки кредитоспроможності потребує вдосконалення.

Доля простроченої заборгованості в загальному об'ємі наданих кредитів



Свідчення про **зниження** рівня прибутковості та рентабельності українських банків

Дані з сайту НБУ

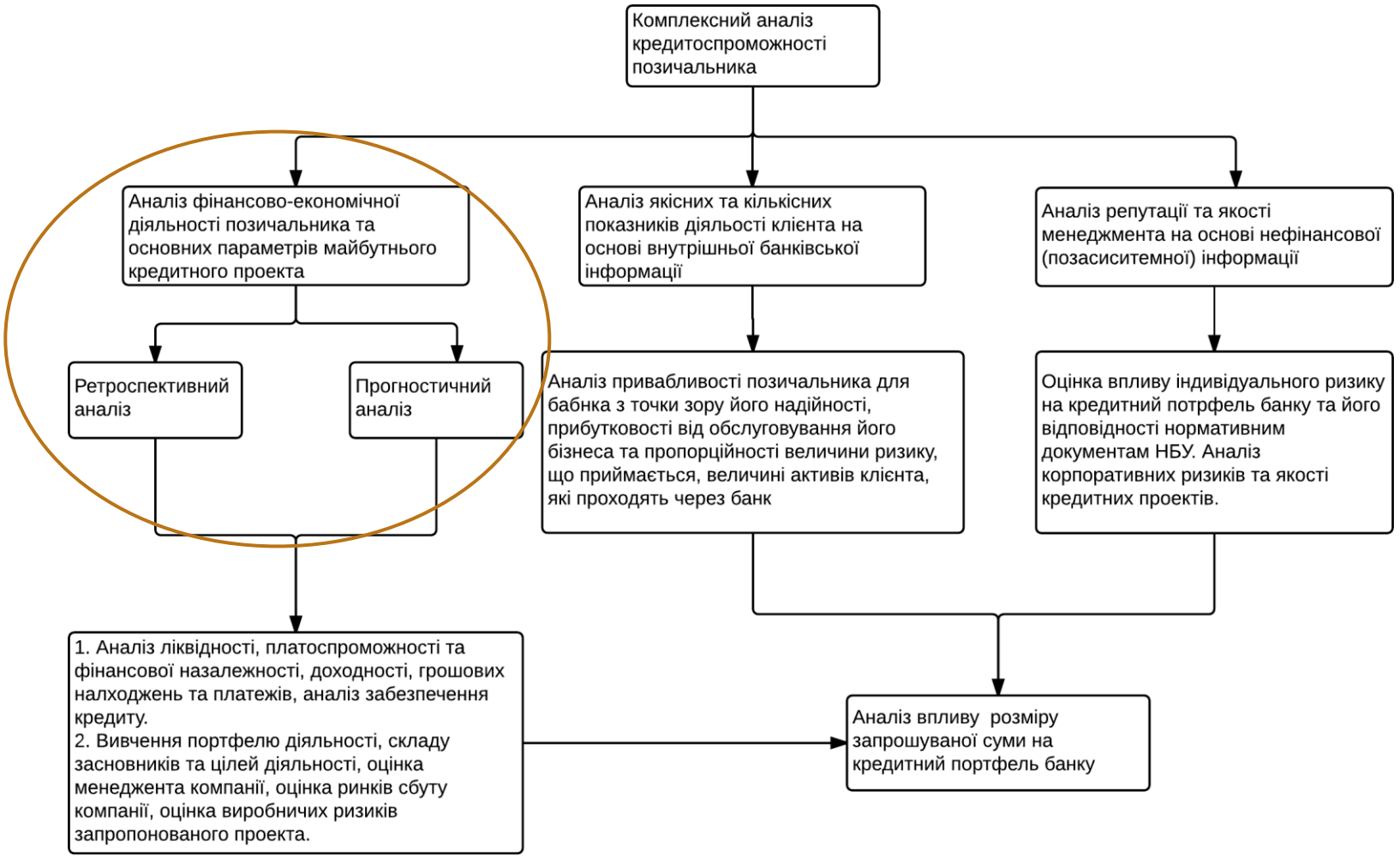
Постановка задачі

- **Кредитний ризик** - ймовірність несплати позичальником основного боргу та відсотків, які належать сплаті за користуванням кредитом у терміни, визначені у кредитному договорі.
- **Кредитоспроможність** визначають з метою оцінки потенційного кредитоотримувача щодо платоспроможності завчасно до прийняття рішення про можливість кредитування та його умов. Метою оцінки кредитоспроможності є попередження або максимальне зменшення кредитного ризику.

Тож, для досягнення поставленої мети **необхідно**:

1. Проаналізувати особливості та обмеження застосування статистичних методів до оцінки кредитоспроможності.
2. Виявили можливості застосування класифікаційних методів у процедурі кредитного скорингу.
3. Визначити шляхи порівняння якості моделей, побудованих з використання методів з п.2.

Процес комплексного аналізу кредитоспроможності позичальника



Підходи до розробки моделей оцінки кредитоспроможності

- статистичні методи
- нейронні мережі
- експертні методи
- логіко-імовірнісні методи
- методи на основі теорії мультимножин
- оптимізаційні методи
- гібридні методи
- методи нечітко-множинного опису

Кредитний скоринг

Методи кредитного скорингу

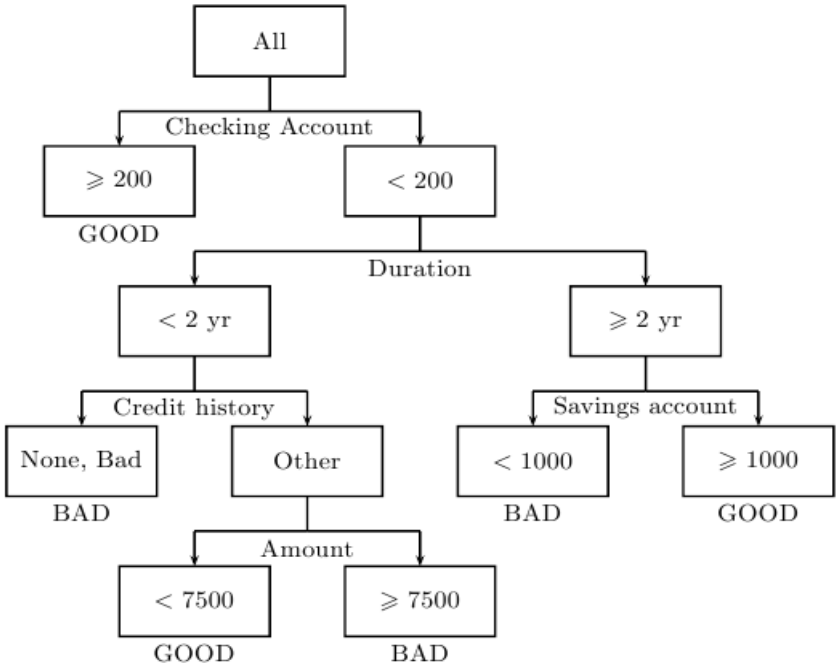
лінійна регресія, логістична регресія, лінійне програмування, дерева класифікації, нейронні мережі, генетичний алгоритм, дискримінантний та інші

Дерева рішень

- **Дерева рішень** використовуються в якості математичних і обчислювальних методів для опису, класифікацыъ і узагальнення набору даних, які записані, наприклад, наступним чином:

$$(x, Y) = (x_1, x_2, x_3, \ldots, x_k, Y),$$

де *Y* – залежна (цільова) змінна



Приклад дерева рішень

Градієнтний бустинг

- Алгоритм градієнтного бустингу дерев рішень будує модель у вигляді суми дерев:

$$f(x) = h_0 + v\sum_{j=1}^{M} h_j(x),$$

де h_0 – початкове наближення,

- Наступна функція призначена для розв'язання задач бінарної класифікації:

$$L(y, y_1', y_2') = -\sum_{K=1}^{2} (y = k) ln(\frac{exp(y_k')}{\sum_{i=1}^{2} exp(y_k')}).$$

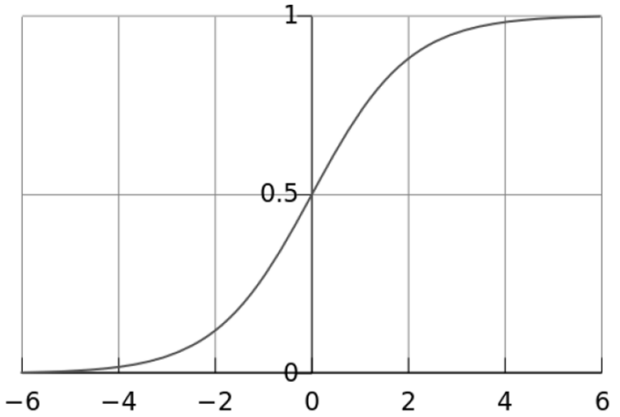
Байєсовський класифікатор

- Наївний байєсовський класифікатор об'єднує модель з правилом рішення. Одне загальне правило має обрати найбільш ймовірну гіпотезу; воно відомо як апостеріорне правило прийняття рішення. Відповідний класифікатор — це функція classify, визначена наступним чином:

$$\operatorname{classify}(f_1, \ldots, f_n) = \arg\max_c p(C = c) \prod_{i=1}^{n} p(F_i = f_i | C = c)$$

Логістична регресія

- Логіт-функція: $f(x) = \frac{1}{1+e^{-x}}$



- Модель: $\mathrm{P(y)} = f(\theta_0 + \theta_1 \mathrm{x}_1 + \theta_2 \mathrm{x}_2 + \cdots + \theta_m \mathrm{x}_\mathrm{m})$

- Оцінювання параметрів моделі (максимізація функції правдоподібності):

$$L^* = \sum_{i \in I_1} ln\, P_i\,(W) + \sum_{i \in I_0} ln\big(1 - P_i\,(W)\big) = \sum_{i=1}^{k} [Y_i\, ln\, P_i\,(W) + (1 - Y_i)\, ln\big(1 - P_i\,(W)\big)]$$

- Задача зводиться до розв’язання системи рівнянь (за методом Ньютона-Рафсона)

$$\begin{cases} g_0(\theta) = \sum_{i=1}^{n} g(\theta^T x^i) - \sum_{\{i:y_i=1\}} 1 = 0, \\ g_i(\theta) = \sum_{i=1}^{n} g(\theta^T x^i)\, x_j^i - \sum_{\{i:y_i=1\}} x_j^i = 0 \;\; (j = 1, \ldots m) \end{cases}$$

Дискримінантний аналіз

- **Дискримінантний аналіз** – метод багатовимірного статистичного аналізу. Він включає в себе методи класифікації багатовимірних спостережень за принципом максимальної подібності за наявності навчальних ознак.

- **Зміст:** на підставі навчальних вибірок перетворити багатовимірний масив на одновимірний показник для прогнозування належності спостережень до визначених груп.

Критерії порівняння якості класифікаційних методів

	Фактично	
Модель	позитивно	негативно
Позитивно	TP	FP
негативно	FN	TN

Таблиця спряженості

Позитивно - повернення кредиту (0)
Негативно – дефолт (1)

FN – помилки другого роду
FP – помилки першого роду

Проста точність (common accuracy, CA):
$$CA = \frac{TP + FP}{N}$$

Точність (Precision, P):
$$P = TPR = \frac{TP}{TP + FN} 100\%$$

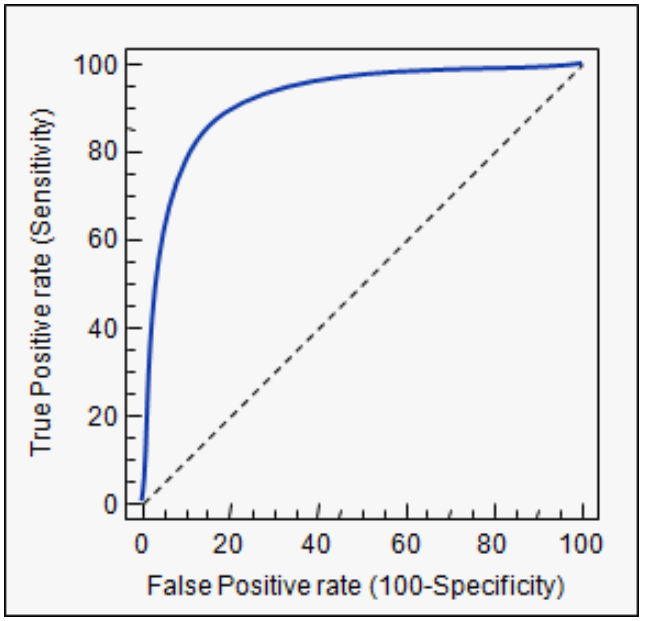
Повнота (Recall, R):
$$R = FPR = \frac{FP}{TN + FP} 100\%$$

Метрика F_1:
$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}$$

Чутливість (Sensitivity)
$$Se = TPR = \frac{TP}{TP + FN} 100\%$$

Специфічність (Specificity)
$$Sp = \frac{TN}{TN + FP} 100\% = 100\% - FPR$$

ROC-крива показує залежність чутливості від специфічності.



Опис вхідних даних

- Джерело: ВАТ «Тинькофф Банк»
- Посилання :

https://static.tcsbank.ru/documents/olymp/SAMPLE_ACCOUNTS.csv

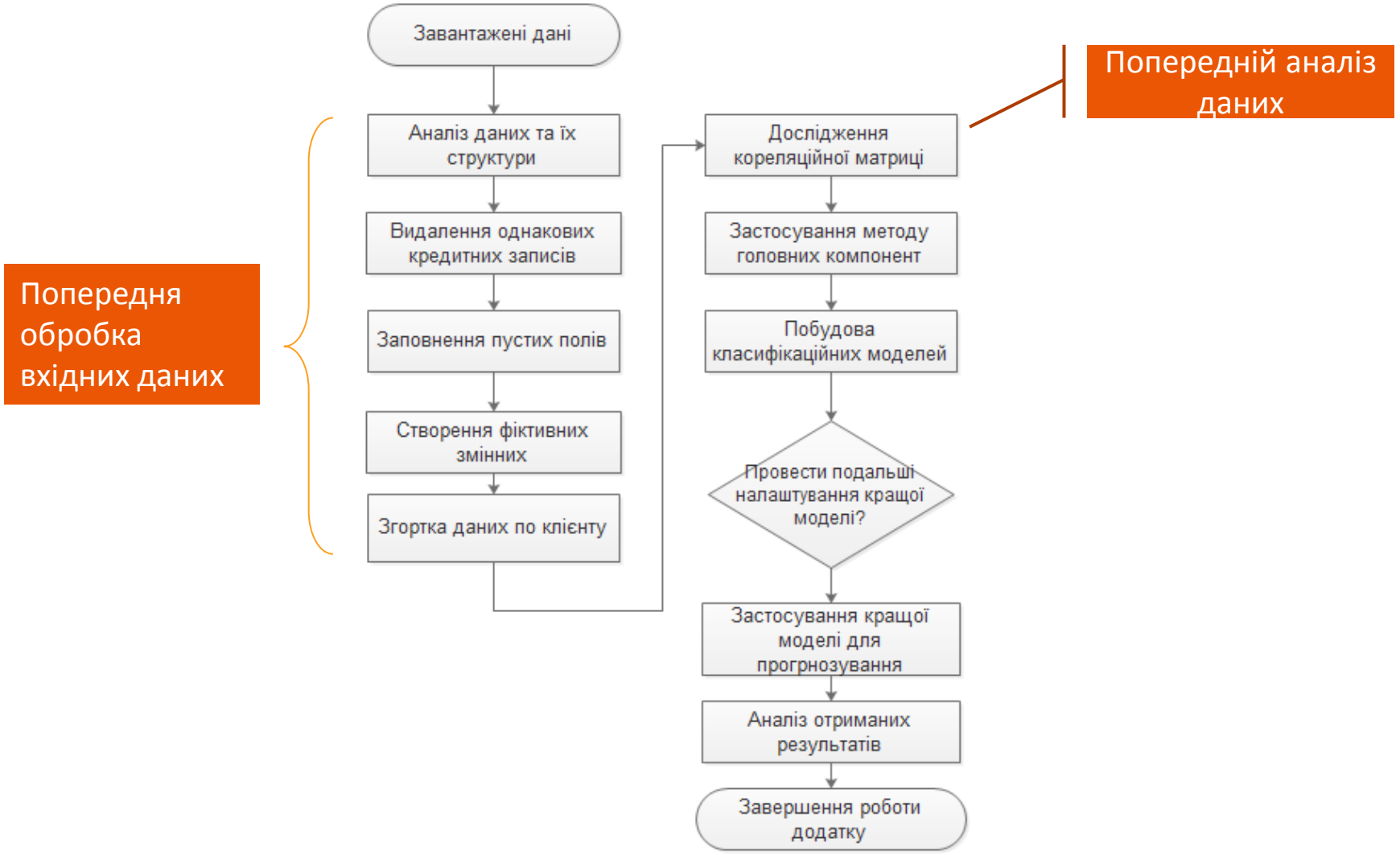
- Розмір вибірки: 280 942 елементи
- Об’єм вибірки: 50 000 клієнтів банку
- Мета: визначити найкращий алгоритм визначення позичальників, схильних до дефолту.

«Дефолт» - допущення клієнтом прострочення на >=90 днів протягом першого року користування кредитом.

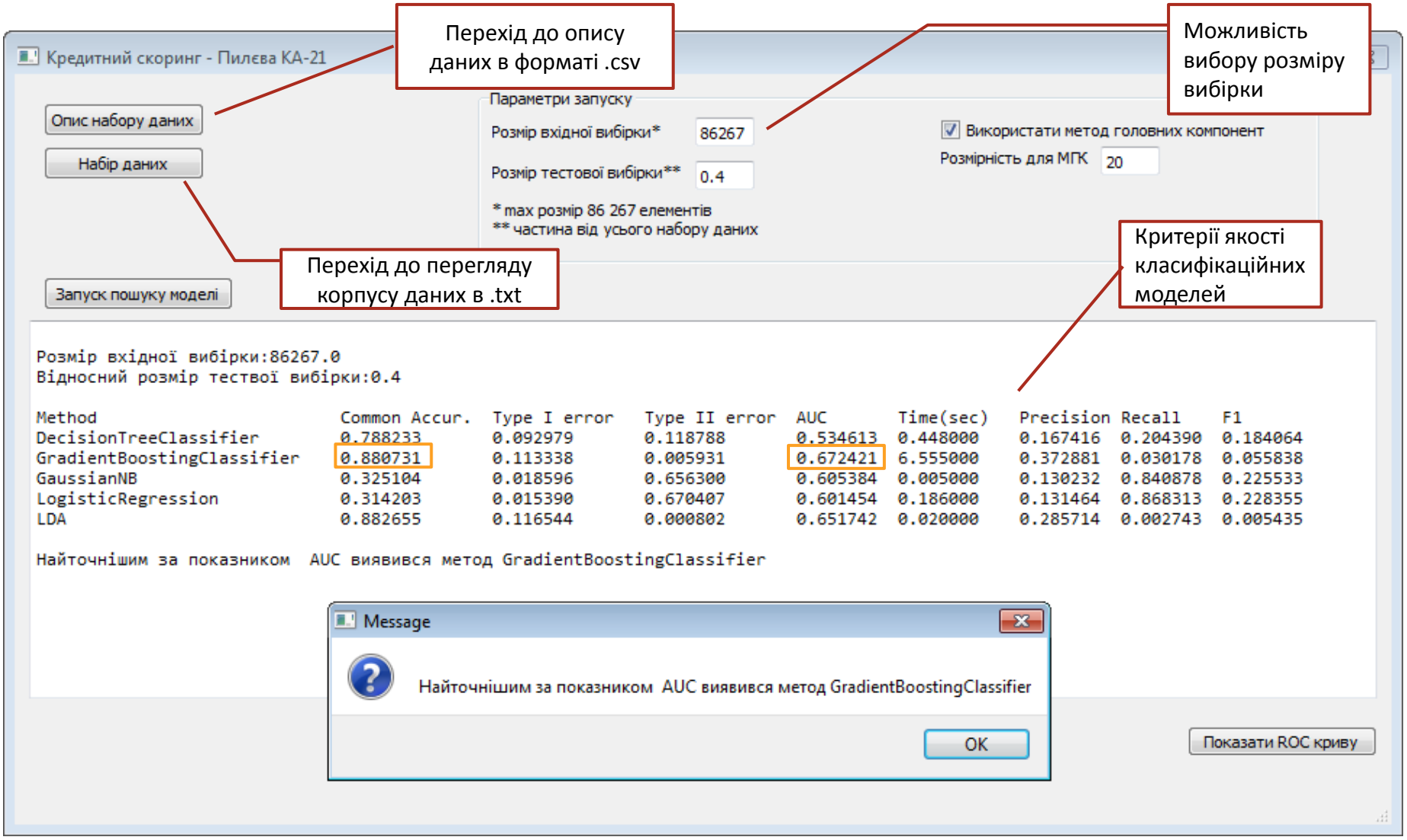
- Опис даних:

Data description

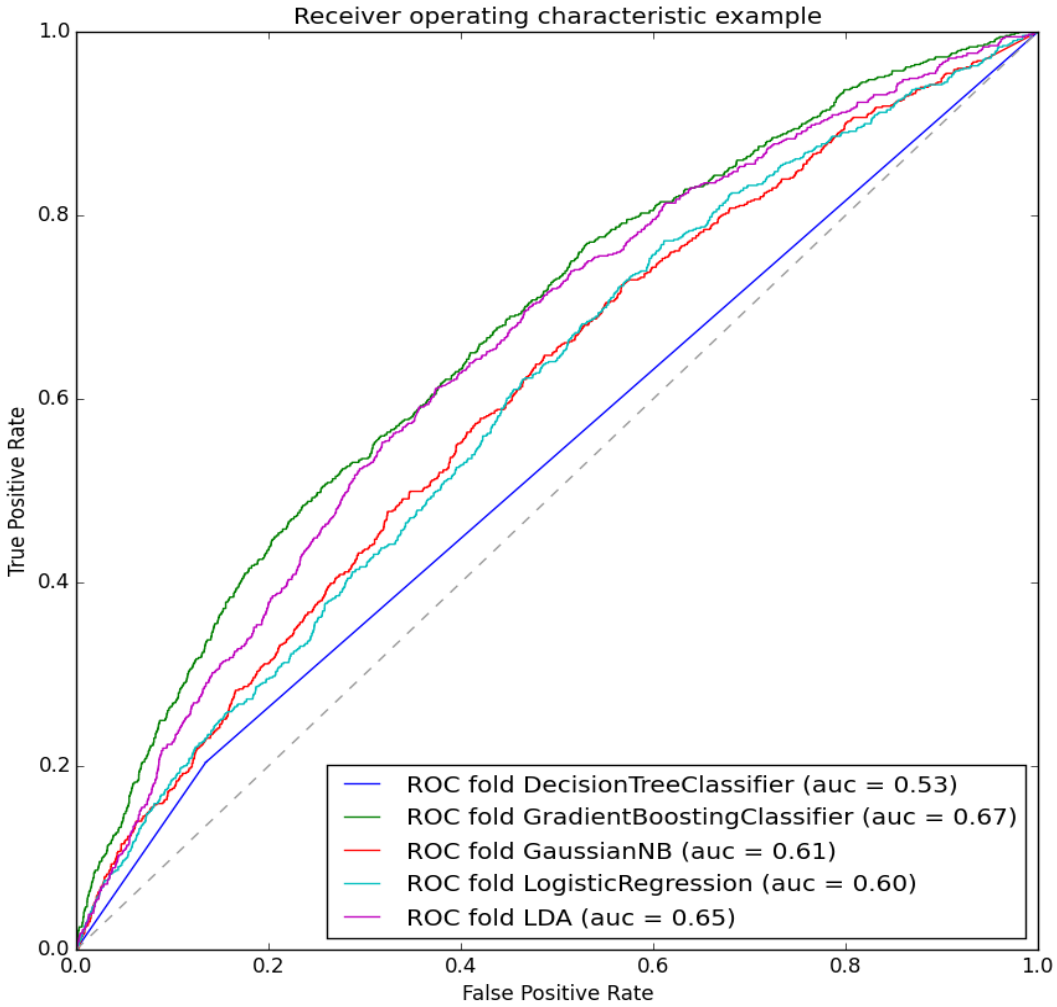
Функціональна схема програмного додатку



Результати роботи скорингової системи



Результати роботи скорингової системи

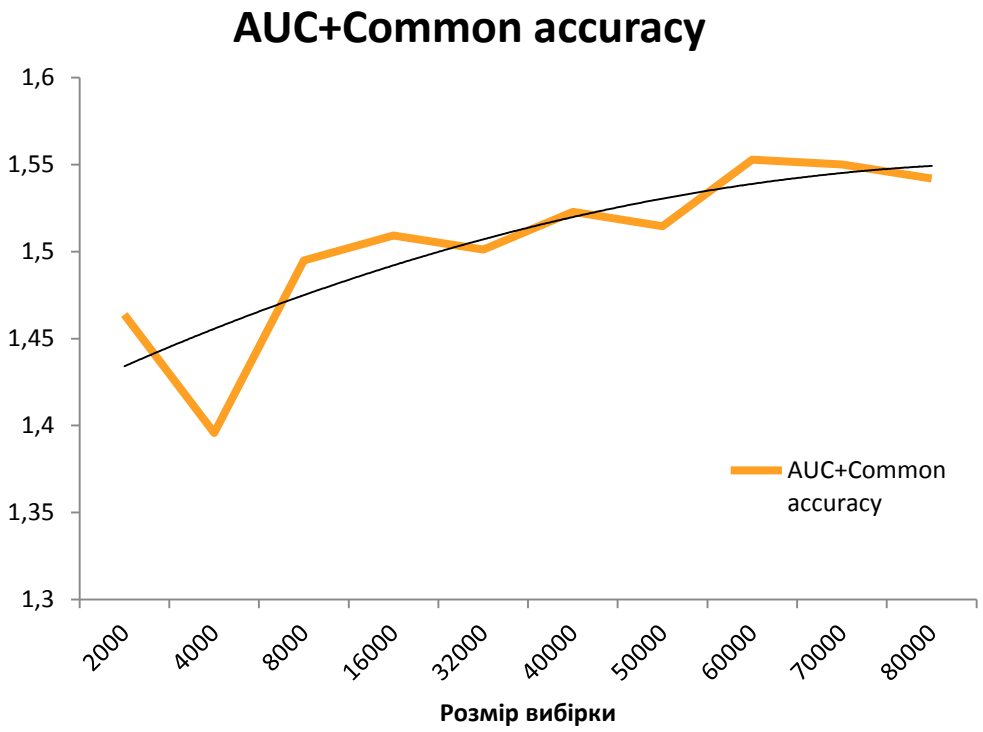


Результати роботи скорингової системи

Розмір вибірки	Краща модель	AUC	Common accuracy	AUC+Common accuracy
2000	LogisticRegression	0,618916	0,844961	1,463877
4000	GradientBoostingClassifier	0,539201	0,856631	1,395832
8000	GradientBoostingClassifier	0,632724	0,862191	1,494915
16000	GradientBoostingClassifier	0,637922	0,871219	1,509141
32000	LogisticRegression	0,621935	0,879145	1,50108
40000	GradientBoostingClassifier	0,649945	0,872867	1,522812
50000	LDA	0,638273	0,876254	1,514527
60000	GradientBoostingClassifier	0,677713	0,875116	1,552829
70000	GradientBoostingClassifier	0,671451	0,878655	1,550106
80000	GradientBoostingClassifier	0,662393	0,87961	1,542003

Результати роботи скорингової системи

Наявна загальна тенденція до зростання якості моделі із збільшенням величини вибірки.



Загальні висновки

- Обґрунтовано необхідність автоматизації процедури кредитного скорингу.
- Обрано множину критеріїв для аналізу якості класифікації позичальників.
- Для оцінки кредитоспроможності на основі даних про кредитні історії позичальників побудовано моделі із застосуванням п’яти методів класифікації та виявлено найточніший з методів на основі зазначених критеріїв.
- Найкращі результати на повній вибірці у 86 267 елементи було отримано з використанням методу градієнтного бустингу, що показав точність 88%.
- Проаналізовано результати роботи побудованого алгоритму та з’ясовано можливості його використання.
- Продемонстровано, що неможливо створити єдиний стандартизований алгоритм оцінювання кредитоспроможності з огляду на різну структуру та специфіку даних кожного комерційного банку.

Перспективи подальшого розвитку

- Створення розширеної скорингової системи із врахуванням особистих даних позичальників та зовнішні умови ринку для поліпшення результатів передбачення дефолтів.
- Розширення набору методів, які застосовуються у системі.
- Використання каскаду класифікаторів.

Питання