НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ

ПРОЦЕСИ КРЕДИТУВАННЯ ФІЗИЧНИХ ОСІБ З ВИКОРИСТАННЯМ НЕЛІНІЙНИХ МОДЕЛЕЙ І БАЙЄСІВСЬКИХ МЕРЕЖ

Студент:
Бойко Оксана Михайлівна
Науковий керівник:
д.т.н., професор Бідюк Петро Іванович

Об’єкт та предмет дослідження

- **Об’єкт дослідження:** Процеси кредитування фізичних осіб, інформаційні технології в аналізі фінансових процесів.
- **Предмет дослідження:** нелінійні математичні моделі і методи байєсівського аналізу даних.

Постановка задачі

- Виконати огляд математичних методів оцінювання кредитних ризиків на основі статистичних даних.
- Виконати аналіз можливості застосування нелінійної регресії та Байєсівських мереж.
- Виконати та проаналізувати результати обчислювальних експериментів з оцінювання кредитоспроможності клієнтів за заданими статистичними даними.
- Виконати порівняльний аналіз результатів застосованих методів.

Оцінювання кредитного ризику. Методи і моделі

- Нечітка логіка
- Лінійні моделі
- Нейронні та нейронечіткі мережі
- Нелінійні моделі
- Байєсівські мережі
- Дерева класифікації
- Банківські методики

У роботі застосовано:

Нелінійні моделі та Байєссівські мережі

Логістична регресія

- Логіт-функція: $f(x) = \frac{1}{1 + e^{-x}}$
- Модель $\Pr = f(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_m x_m)$
- Оцінювання параметрів моделі

(максимізація функції правдоподібності)

$$\log L(\theta) = \sum_{i=1}^{m} \log \mathbb{P}\{y = y^{(i)} \mid x = x^{(i)}\} = \sum_{i=1}^{m} y^{(i)} \log f(\theta^T x^{(i)}) + (1 - y^{(i)}) \log(1 - f(\theta^T x^{(i)})).$$

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

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

Мережі Байєса

- Поєднання графічного представлення певного процесу з його ймовірнісним характером;
- можливість аналізу варіантів розвитку процесу;
- відстеження правильності встановлення причинно-наслідкового зв'язку між окремими факторами;
- дозволяє підвищити обґрунтованість рішень при аналізі складних проблемних ситуацій.

Мережа Байєса - це трійка $N = < V, G, J >$,

де G — множина змінних;

J — спрямований ациклічний граф, вузли якого відповідають випадковим змінним модельованого процесу;

$V = \{X_1, X_2, ..., X_n\}$ — спільний розподіл ймовірностей змінних

.

Методика побудови мережі Байєса

1. Виконати аналіз проблеми і зробити формалізовану постановку задачі; сформулювати питання, на яке має бути отримана ймовірнісна відповідь в результаті формування ймовірнісного висновку за допомогою побудованої мережі.

2. Визначити множину даних, що характеризують змінні задачі, отримати їх експертні оцінки та/або статистичні дані.

3. Поставити у відповідність усім отриманим даним взаємовиключні змінні.

Методика побудови мережі Байєса (продовження)

4. Побудувати ациклічний граф, що відображає істотні умови незалежності змінних та існування причинно-наслідкових зв’язків.

5. Визначити апріорні ймовірності та оптимізувати топологію мережі на основі наявної інформації.

6. Виконати навчання мережі і сформувати висновок по відношенню до відповідних станів процесу.

7. Проаналізувати отримані результати і зробити висновки щодо ймовірності очікуваної події.

Критерії якості моделей

- **Сума квадратів залишків**

$$\sum_{i=1}^{n}(y_i - \hat{y})^2$$

- **Критерій Ханна-Куїна**

$$HQ = -2\frac{lnL}{n} + 2 * k * \frac{\ln(lnn)}{n},$$

де L - функція правдоподібності, n - кількість спостережень, k - кількість змінних в моделі.

- **Критерій Макфадена**

$$McFaddenRsquared = 1 - \frac{lnL}{RestrlnL},$$

де lnL – логарифм функції правдоподібності, $RestrlnL$ – залишок логарифма функції правдоподібності.

Помилки I-го та II-го роду

	Прогноз моделі: Повернення кредиту (0)	Прогноз моделі: Дефолт (1)
Фактично: Повернення кредиту (0)	Вірно класифіковані (TP)	Помилки II-го роду (FN)
Фактично: Дефолт (1)	Помилки I-го роду (FP)	Вірно класифіковані (TN)

Відносні показники у процентах:

- частка істинно позитивних прикладів (True Positives Rate):

$$TPR = \frac{TP}{TP + FN}$$

- частка хибно позитивних прикладів (False Positives Rate):

$$FPR = \frac{FP}{TN + FP}$$

Чутливість моделі – це частка істинно позитивних випадків, тобто

$$Se = TPR = \frac{TP}{TP + FN}$$

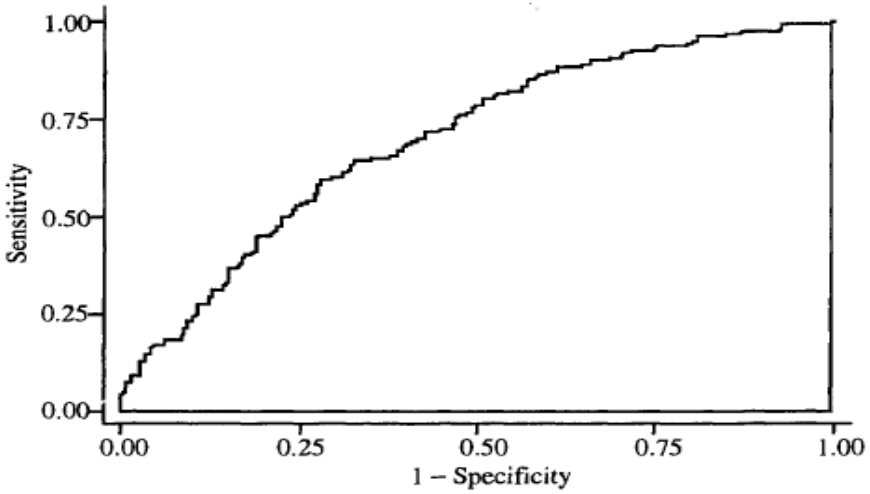
Специфічність моделі – це частка істинно негативних випадків, які були вірно класифіковані моделлю:

$$Sp = \frac{TN}{TN + FP}$$

$$Sp = \frac{TN + FP - FP}{TN + FP} = 1 - \frac{FP}{TN + FP} = 1 - FPR$$

ROC-крива та індекс GINI

ROC-крива показує залежність кількості вірно класифікованих позитивних прикладів від кількості невірно класифікованих негативних прикладів.



Індекс GINI – це площа області між діагоналлю і кривою Лоренца, поділена на площу усієї області під діагоналлю. Індекс GINI широко використовується для аналізу роздільної здатності системи оцінювання при управлінні кредитними ризиками, тобто оцінки здатності моделі розділяти клієнтів на схильних та несхильних до дефолту.

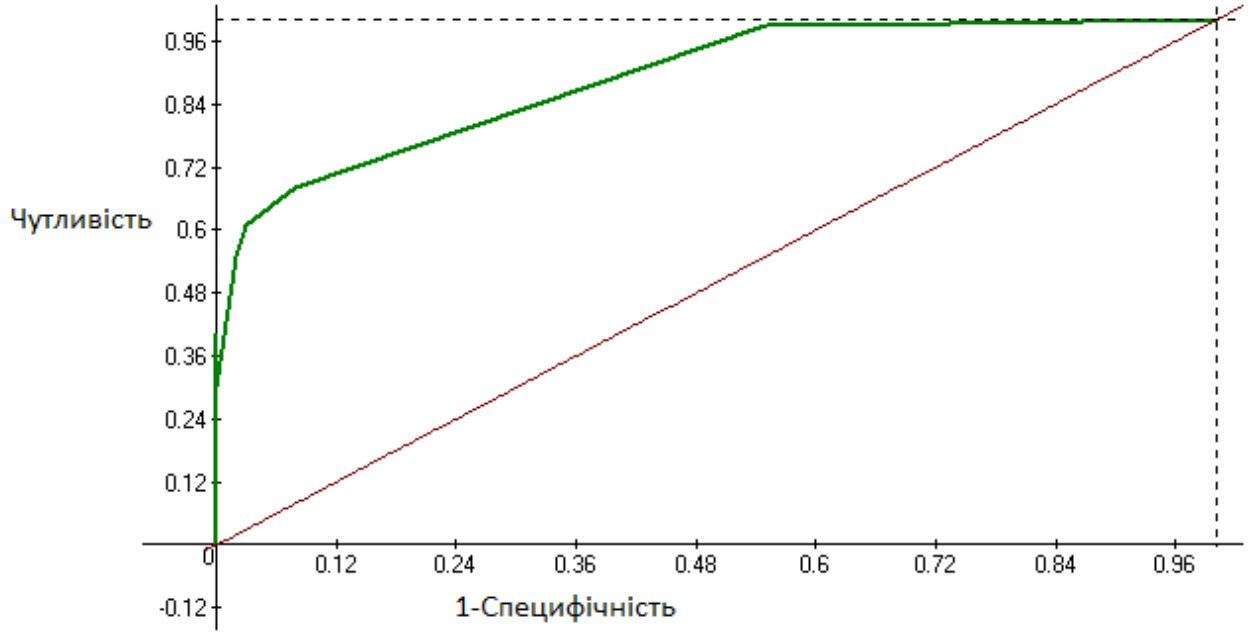
Результати застосування нелінійної регресії

Варіант розрахунків	Якість моделі			Якість прогнозу			
	Сума квадратів залишок	Критерій Хана-Куїна	Критерій Макфадена	За ймовірністю прийняття		За обчислення математич. сподівання	
				0	1	0	1
All_1	708.0545	0.394975	0.012778	0.00	100.00	5.54	95.03
BASE_ALL	347.9989	0.382673	0.038641	0.00	100.00	6.48	95.17
BASE_743	106.1273	0.921421	0.372007	76.01	84.41	71.52	71.60
NK_ALL	249.4353	1.003179	0.193599	91.30	42.89	78.22	21.78
NK_588	104.7132	0.179674	0.230992	71.43	76.45	64.39	64.27

Загальна точність моделі та помилки I-го і II-го роду для різних рівнів порогу відсікання, отримані для логістичної регресії (вибірка ALL)

	Прогноз: Повернення кредиту (0)	**Прогноз: Дефолт (1)**	**Процент точності**
Cut-off=0,5			
Факт: Повернення кредиту (0)	90	42	0.68
Факт: Дефолт (1)	5	63	0.92
Загальна точність моделі			**0.77**
Cut-off=0,4			
Факт: Повернення кредиту (0)	87	54	0,61
Факт: Дефолт (1)	2	57	0,97
Загальна точність моделі			**0,72**
Cut-off=0,3			
Факт: Повернення кредиту (0)	83	67	0.55
Факт: Дефолт (1)	1	49	0.98
Загальна точність моделі			**0.66**

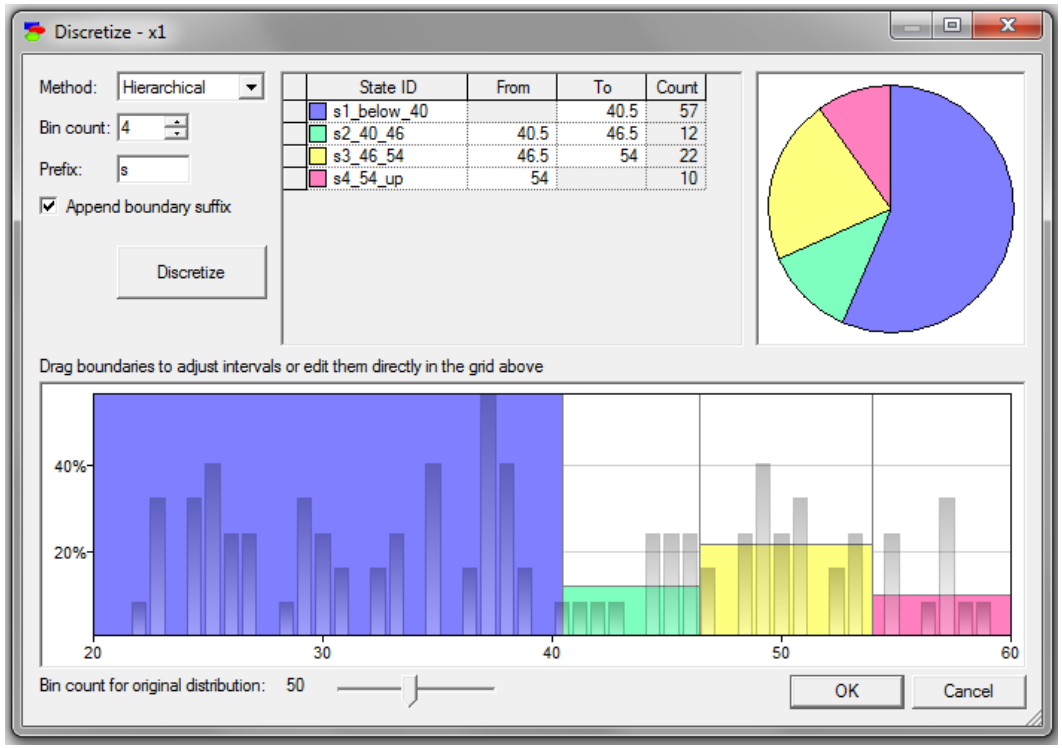
ROC-крива для бінарної логістичної регресії



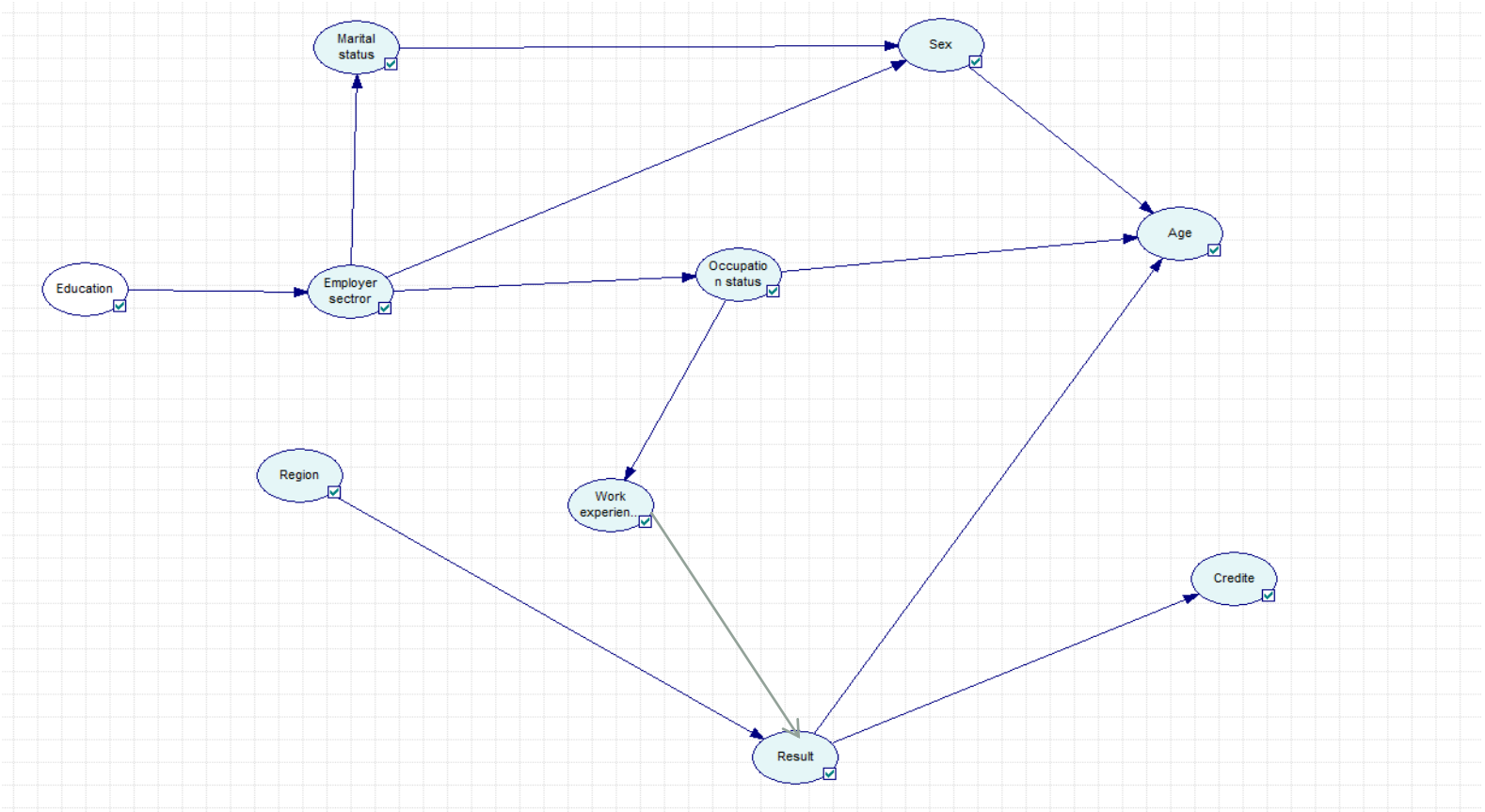
$$GINI = 2 * AUC - 1 = 0,74$$

Побудова моделей у формі Байєсівських мереж

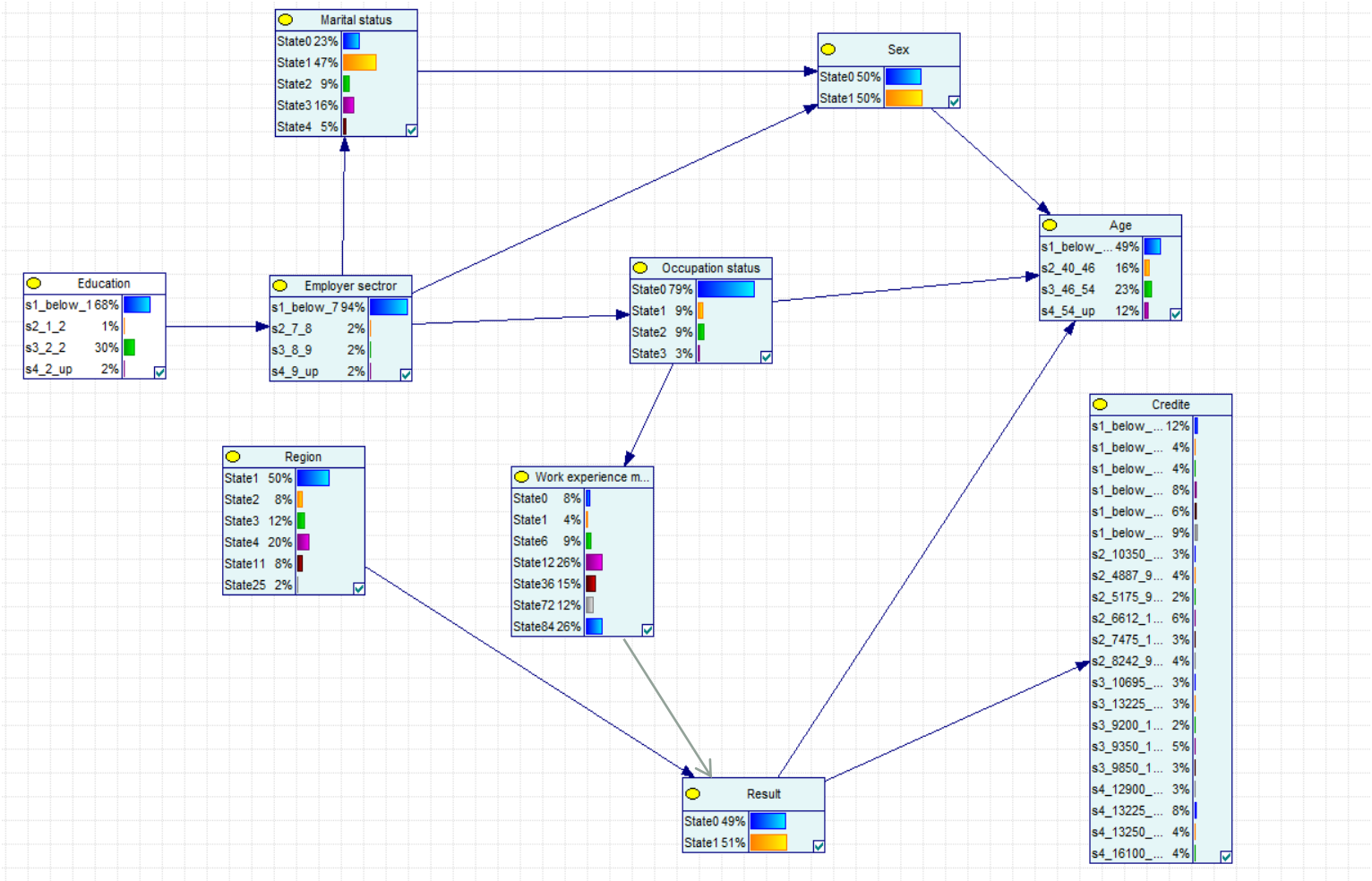
- Дискретизація змінних



Структура мережі Байєса: вибірка BASE



Структура мережа Байєса у вигляді гістограм вибірки BASE



Загальна точність моделі та помилки I-го і II-го роду для різних рівнів порогу відсікання - мережі Байєса (BASE)

	Прогноз: Повернення кредиту (0)	Прогноз: Дефолт (1)	Процент точності
Cut-off=0,5			
Факт: Повернення кредиту (0)	100	49	0.67
Факт: Дефолт (1)	11	40	0.78
Загальна точність моделі			**0.7**
Cut-off=0,4			
Факт: Повернення кредиту (0)	105	45	0.7
Факт: Дефолт (1)	7	43	0.86
Загальна точність моделі			**0.74**
Cut-off=0,3			
Факт: Повернення кредиту (0)	105	43	0.709
Факт: Дефолт (1)	6	44	0.88
Загальна точність моделі			**0.745**

Порівняльна таблиця характеристик для моделей Байєса та бінарної логістичної регресії

All				
Назва методу	**Індекс GINI**	**Значення AUC**	**Точність моделі**	**Якість моделі**
Бінарна логістична регресія	**0,74**	0,87	**0,77**	Дуже висока
Мережа Байєса	0,72	0,86	0,74	Дуже висока
NK				
Назва методу	**Індекс GINI**	**Значення AUC**	**Точність моделі**	**Якість моделі**
Бінарна логістична регресія	0,67	0,83	0,87	Дуже висока
Мережа Байєса	**0,7**	0,85	0,87	Дуже висока
BASE				
Назва методу	**Індекс GINI**	**Значення AUC**	**Точність моделі**	**Якість моделі**
Бінарна логістична регресія	0,74	0,87	**0,79**	Дуже висока
Мережа Байєса	**0,76**	0,88	0,74	Дуже висока

Перспектива подальших досліджень

- Розробка комбінованих математичних моделей для аналізу кредитоспроможності позичальників з використанням регресійних та байєсівських моделей. Обчислення усереднених оцінок прогнозів.
- Застосування комбінованого оптимізаційного підходу до обчислення параметрів моделей: наприклад, ММП + Монте-Карло для марковських ланцюгів.
- Розробка та програмна реалізація СППР для аналізу кредитних ризиків на основі кількох типів математичних моделей та критеріїв якості даних, моделей та оцінок прогнозів.

Висновки

- Виконано короткий огляд сучасних систем для оцінювання фінансових ризиків. Проаналізовано їх переваги та недоліки.
- Докладно описано методи, які можна використати для прогнозування кредитоспроможності фізичних осіб: логістична регресія та Байєсівські мережі.
- За допомогою програми GeNIe 2.0 і програми EViews 7.0 побудовано мережі Байєса та нелінійні логістичні регресії. Сформульовано рекомендації для їх практичного застосування.

Висновки(продовження)

- Встановлено, що кращі результати класифікації клієнтів у даному випадку отримано за допомогою мережі Байєса (Індекс GINI=0,76, Значення AUC=0,88).

Наукова новизна отриманих результатів

- На основі статистичних даних стосовно позичальників кредитів побудовані нелінійні математичні моделі типу логіт для класифікації кредитоспроможності позичальників, які забезпечують високу якість класифікації.
- Запропоновано ймовірнісні моделі у формі Байєсівських мереж для прогнозування кредитоспроможності клієнтів, які використано для порівняння з нелінійною регресією.
- Удосконалена методика побудови Байєсівських мереж, яка забезпечує підвищення якості моделі і точності класифікації.

Публікації за результатами дослідження

1. Бойко О.М. «Аналіз кредитних ризиків регресійними методами» // Системні науки і кібернетика (№1) 2016
2. Бойко О. М. «Моделювання кредитоспроможності фізичних осіб»// Системний аналіз та інформаційні технології: матеріали 18-ї Міжнародної науково-технічної конференції SAIT 2016, Київ/ ННК “ІПСА” НТУУ “КПІ”. – К.: ННК “ІПСА” НТУУ “КПІ”, 2016. – 304 с.

Дякую за увагу!