



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

Тема: Розробка програмного продукту для оцінки кредитоспроможності фізичних осіб з використанням методів інтелектуального аналізу даних

Виконавець роботи:
студентка 4 курсу
групи КА-24
Бакун Сабіна Антонівна

Науковий керівник:
к.т.н., старший викладач
каф. ММСА Терентьєв
Олександр Миколайович

- Мета дослідження
 - полягає у вивченні методів побудови прогнозуючих моделей та їх використанні для оцінювання кредитних ризиків фінансових установ
- Об'єкт дослідження
 - набір статистичних даних щодо наданих банком споживчих кредитів фізичним особам
- Предмет дослідження
 - лінійна імовірнісна модель, логістична регресія, дерева рішень

Актуальність дослідження

- Останні роки в Україні спостерігається збільшення банківського споживчого кредитування. Не дивлячись на те, що споживче кредитування є одним з найбільш прибуткових видів банківської діяльності, воно водночас є найбільш ризикованим, оскільки багато банків зіштовхуються з проблемою неповернення виданих фізичним особам кредитів.
- У зв'язку з цим застосування та розробка нових більш досконалих методик оцінювання кредитоспроможності осіб і ризику банків при наданні кредитів в умовах сучасних кризових явищ у фінансовій сфері є критичним.

Постановка задачі

- Проаналізувати існуючі рішення для бізнес аналітики у сфері кредитних ризиків.
- Виконати огляд і аналіз існуючих математичних методів моделювання і прогнозування кредитних ризиків.
- Розробка архітектури системи підтримки прийняття рішень для аналізу, моделювання та прогнозування ймовірності повернення наданих споживчих кредитів.
- Апробація запропонованих методів у вигляді комп'ютерної програми на реальних статистичних даних.
- Аналіз та порівняння отриманих результатів моделювання та прогнозування.

Кредитоспроможність та кредитний скоринг

- Кредитоспроможність — це здатність позичальника повністю та у визначений строк розрахуватися за своїми борговими зобов'язаннями.
- Кредитний скоринг — це система оцінки кредитоспроможності, що на основі кредитних історій банку оцінює ймовірність дефолту потенційного позичальника, виходячи з його соціально-демографічних характеристик.



Credit Score

- ☒ Excellent
- ☐ Average
- ☐ Poor

Ринок платформ бізнес-аналітики

- Головні вендери, що розробляють системи ризик менеджменту:



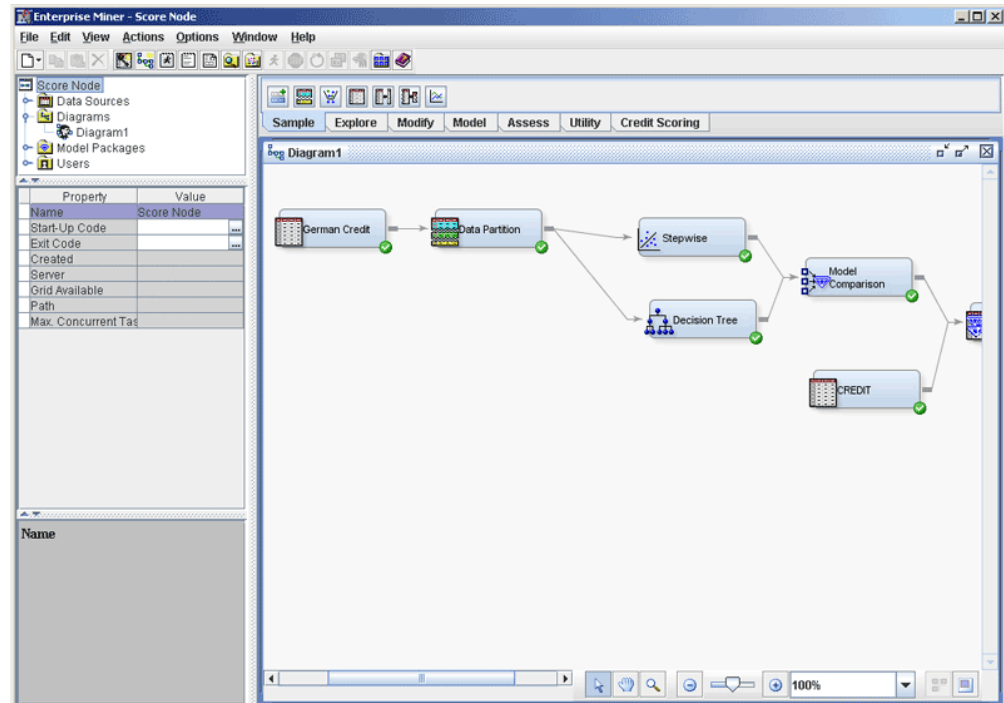
Магічний квадрант платформ бізнес-аналітики (Gartner Magic Quadrant for Advanced Analytics Platforms 2016)

- Магічний квадрант Gartner - це рейтинг, побудований у вигляді системи координат, де на горизонтальній осі вимірюється повнота бачення (completeness of vision), а на вертикальній осі - здатність реалізації (ability to execute).
- Лідерами стають постачальники з позитивними оцінками за обома показниками (верхній правий квадрант).
- За результатами оцінювання у 2016 році лідерами на ринку бізнес-аналітики є компанії Dell, SAS, IBM, KNIME і RapidMiner.



Огляд SAS® Enterprise Miner™

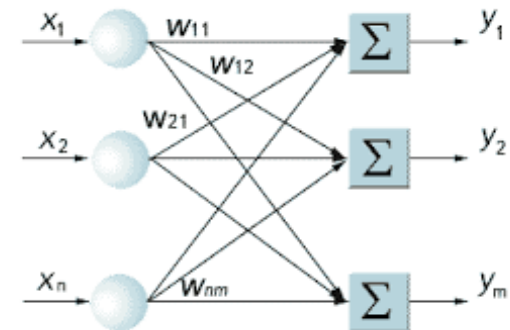
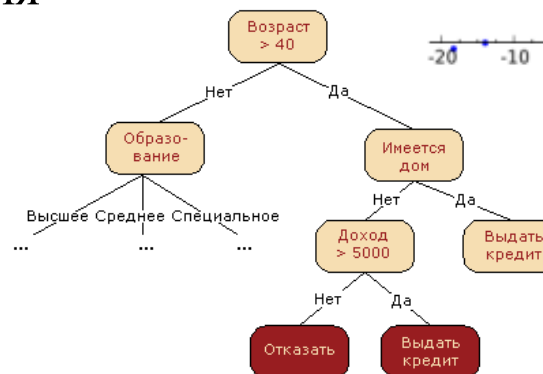
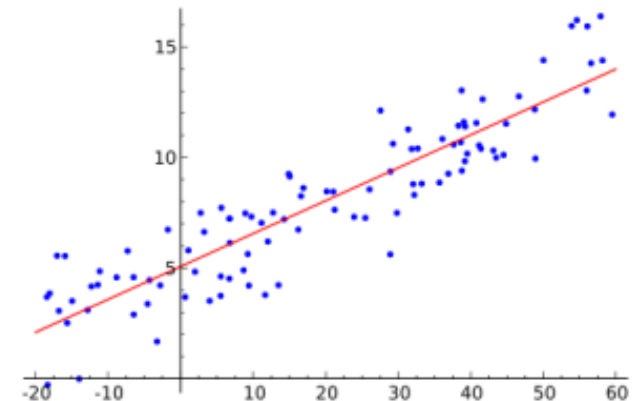
- SAS Enterprise Miner пропонує багатий і простий у використанні набір методів вилучення з даних цінної інформації, що допомагає в прийнятті рішень.
- Програмне забезпечення SAS для інтелектуального аналізу даних застосовується при вирішенні таких завдань, як виявлення випадків шахрайства, мінімізація ризиків, оцінка і прогнозування потреб в ресурсах, зниження відтоку клієнтів і багатьох інших.



Математичні методи для побудови моделей прогнозування та оцінки кредитного ризику

Використовуються наступні методи:

- Регресійний аналіз:
 - Лінійна регресія
 - Логістична регресія
- Древа рішень
- Мережі Байєса
- Нейронні мережі
- Інші методи



Лінійна імовірнісна модель

- Лінійна імовірнісна модель - це модель у формі лінійної регресії, залежна змінна якої набуває значення 0 або 1 залежно від того, яким є результат повернення особою наданого споживчого кредиту.
- Формально така модель має вигляд:

$$y(k) = b_0 + b_1x_1(k) + b_2x_2(k) + \dots + b_mx_m(k) + e(k),$$

де y – залежна змінна;

$b_0, \dots, b_m$ – параметри рівняння регресії;

$x_1, \dots, x_m$ – пояснювальні змінні, (характеристики клієнта);

$e(k)$ – випадковий процес, зумовлений наявністю невимірюваних збурень, а також помилок оцінювання параметрів моделі;

k – ідентифікатор клієнта.

Метод найменших квадратів

- Для оцінювання невідомого вектора параметрів лінійної регресії використовують метод найменших квадратів (МНК). Відповідно до цього методу мінімізується сума квадратів відхилень, або функцію втрат Q :

$$\begin{aligned} Q(\theta) &= \sum_{k=1}^n e(k)^2 = \sum_{k=1}^n (y(k) - \hat{b}_0 - \hat{b}_1 x_1(k) - \hat{b}_2 x_2(k) - \dots - \hat{b}_m x_m(k))^2 = \\ &= (\vec{y} - X\vec{\theta})^T (\vec{y} - X\vec{\theta}) = \vec{y}^T \vec{y} - 2\vec{y}^T X\vec{\theta} + \vec{\theta}^T X^T X\vec{\theta} \Rightarrow \min \end{aligned}$$

- Оцінкою методу найменших квадратів є вектор $\hat{\theta}$, що мінімізує суму квадратів відхилень:

$$\hat{\theta} = (X^T X)^{-1} X^T \vec{y}.$$

Метод найменших квадратів з обрамленням

- Модифікацією МНК є рекурентний метод розв'язання систем лінійних рівнянь, а саме метод найменших квадратів з обрамленням.
- Процедура МНКО полягає у вираженні вектора оцінок параметрів регресії $\hat{\theta}_{s+1}$ через $\hat{\theta}_s$:

$$\hat{\theta}_{s+1} = \begin{bmatrix} \hat{\theta}_s + \Delta \hat{\theta}_s \\ \hat{\vartheta}_{s+1} \end{bmatrix}, s = \overline{0, m},$$

де $\hat{\vartheta}_{s+1}$ — коефіцієнт при (s+1)-му регресорі, введеному в модель;
m — загальна кількість регресорів.

- Алгоритм МНКО:

$$1) \bar{a}_{s+1} = H_s^{-1} \bar{h}_{s+1}, \bar{h}_{s+1} = X_s^T \bar{x}_{s+1}, H_0^{-1} = 0;$$

$$2) \beta_{s+1} = \eta_{s+1} - \bar{h}_{s+1}^T \bar{a}_{s+1}, \eta_{s+1} = \bar{x}_{s+1}^T \bar{x}_{s+1};$$

$$3) \hat{\vartheta}_{s+1} = \frac{1}{\beta_{s+1}} (\gamma_{s+1} - \bar{h}_{s+1}^T \hat{\theta}_s), \gamma_{s+1} = \bar{x}_{s+1}^T \bar{y};$$

$$4) \Delta \hat{\theta}_s = -\hat{\vartheta}_{s+1} \bar{a}_{s+1}, \Delta \hat{\theta}_0 = 0;$$

$$5) H_{s+1}^{-1} = \begin{bmatrix} H_s^{-1} + \frac{1}{\beta_{s+1}} \bar{a}_{s+1} \bar{a}_{s+1}^T & -\frac{1}{\beta_{s+1}} \bar{a}_{s+1} \\ -\frac{1}{\beta_{s+1}} \bar{a}_{s+1}^T & \frac{1}{\beta_{s+1}} \end{bmatrix}.$$

Оцінка значущості змінних

- Показник інформаційного значення або IV (Information Value):

$$IV = \sum_{i=1}^k (d_i^{(1)} - d_i^{(2)}) * \ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right),$$

де $d_i^{(1)}$ і $d_i^{(2)}$ – відносні частки «хороших» та «поганих» кредитів в i -й категорії, k – кількість категорій незалежної змінної.

Значення IV	Прогнозна здатність
менше 0.02	Не має
0.02-0.1	Низька
0.1-0.3	Середня
0.3-0.5	Висока
більше 0.5	Чудова

Перетворення категоріальних змінних в числові

- Способи трансформації категоріальних змінних в числові:

- 1) порядкова нумерація категорій;
- 2) використання значень коефіцієнту WOE (Weight of evidence), який показує міру інформативності відповідного класу змінної, та обчислюється за формулою:

$$WOE = \ln\left(\frac{d_i^{(1)}}{d_i^{(2)}}\right),$$

де $d_i^{(1)}$ і $d_i^{(2)}$ – відносні частки «хороших» та «поганих» кредитів.

- Приклад перетворення змінної “Ціль кредиту”:

1)

Номер категорії	Назва категорії
1	Автомобіль
2	Відпочинок
3	Меблі
4	Побутова техніка
5	Покупки в магазинах

2)

Значення WOE	Назва категорії
-0.21	Автомобіль
-0.11	Побутова техніка
0	Відпочинок
0.05	Покупки в магазинах
0.19	Меблі

Статистичні характеристики для порівняння моделей

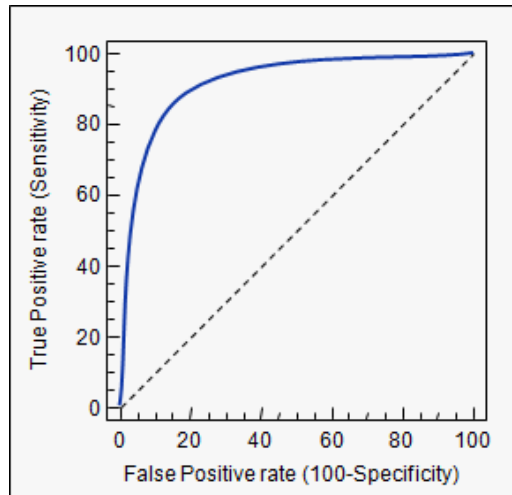
- Загальна точність моделі (CA – англ. Common Accuracy) – обчислюється як відношення вірно спрогнозованих значень до загальної кількості значень N :

$$CA = \frac{\text{кількість вірно спрогнозованих значень}}{N}.$$

- Помилки першого і другого роду:

Результат класифікації моделі	Фактична приналежність	
	Позитивний	Негативний
Позитивний	TP - Істинно позитивні	FP - Хибно позитивні (Помилка 2го роду)
Негативний	FN - Хибно негативні (Помилка 1го роду)	TN - Істинно негативні

Статистичні характеристики для порівняння моделей



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

- Індекс GINI – переводить значення площі під ROC-кривою в діапазон від 0 до 1. Розраховується за формулою:

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

де AUC – площа під ROC-кривою.

Індекс GINI	Якість моделі
0.8 – 1.0	Відмінна
0.6 – 0.8	Дуже висока
0.4 – 0.6	Прийнятна
0.2 – 0.4	Середня
0 – 0.2	Незадовільна

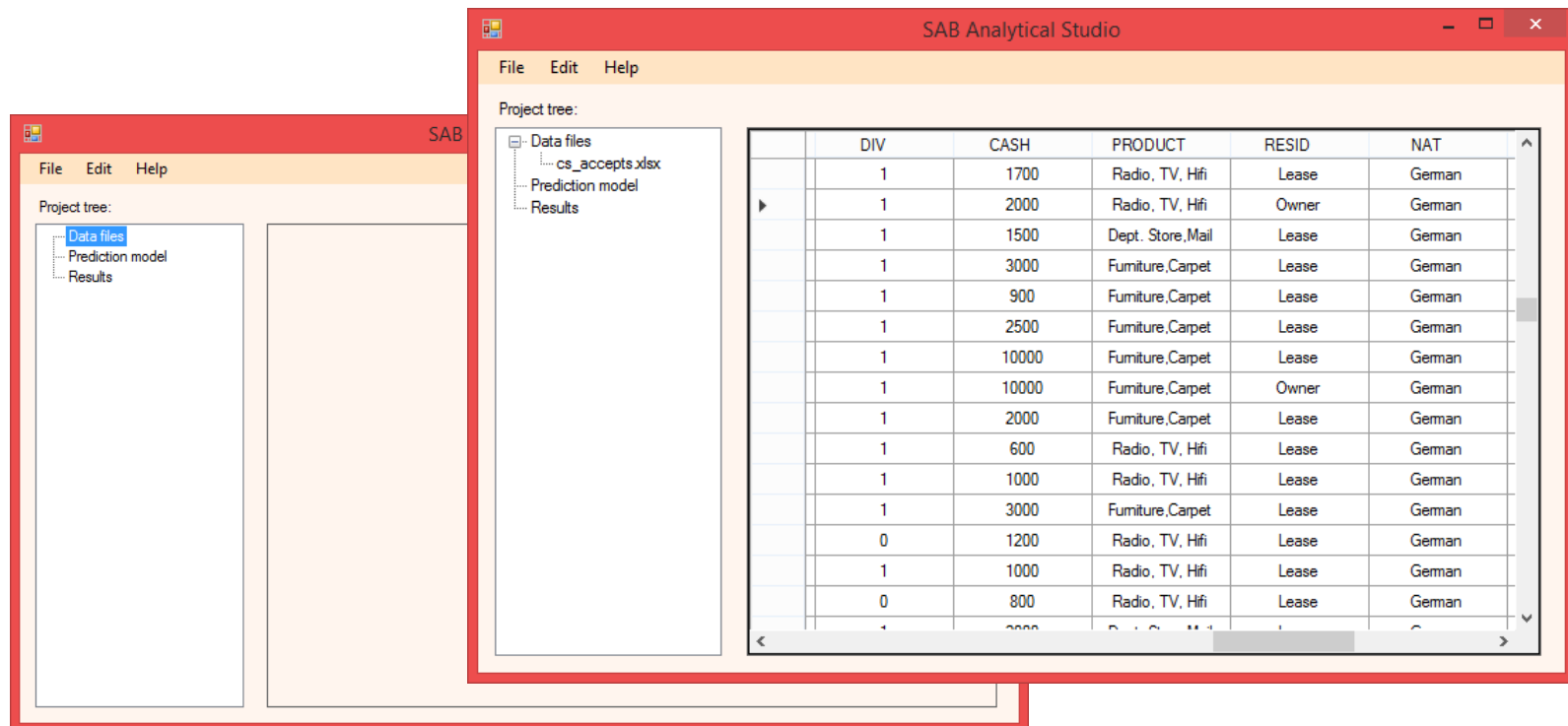
Блок-схема комплексної системи оцінювання і прогнозування кредитних ризиків



Розроблений програмний продукт

- На базі запропонованої архітектури СППР за допомогою мови програмування C# розроблено програмний продукт SAB Analytical Studio для оцінювання кредитоспроможності фізичних осіб на основі методу лінійної імовірнісної моделі.

Головне вікно програми:



Розроблений програмний продукт

Вікно завантаження даних:

Import data file

Data source

Path:

Type of data file:

☒ The first line contains field names

Вікно заповнення пропусків:

Missing value

Columns:

☒ Replace with special value

☐ Replace with average value

Вікно дискретизації даних:

Discretization

Columns:

Methods:

Count:



Group	From	To	Count
1g (- - 37)	-	37	10364
2g (37 - 45)	37	45	10345
3g (45 - 51)	45	51	9813
4g (51 - 57)	51	57	9333
5g (57 - 63)	57	63	9008
6g (63 - 71)	63	71	8927
7g (71 - =)	71	=	7745

Вікно побудови моделі:

Build prediction model (step 2)-cs_accepts.xlsx

Select columns:

Status	Name	IV
☐	TITLE	0.0046
☒	CHILDREN	0.176
☒	AGE	1.9895
☒	NMBLOAN	0.368
☐	FINLOAN	0.0162
☒	INCOME	0.7969

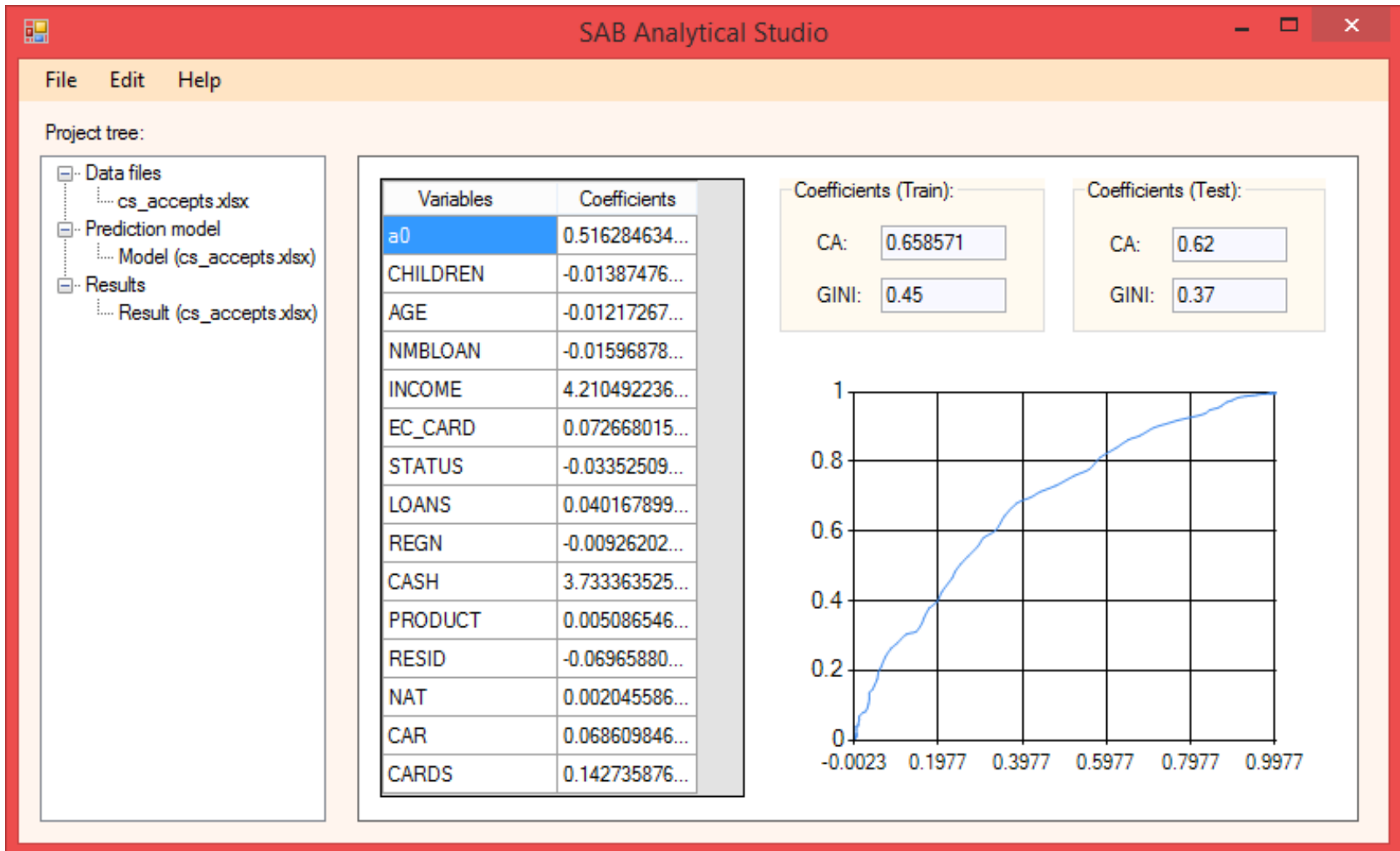
☒ Use WOE for categorical variables

Опис набору даних

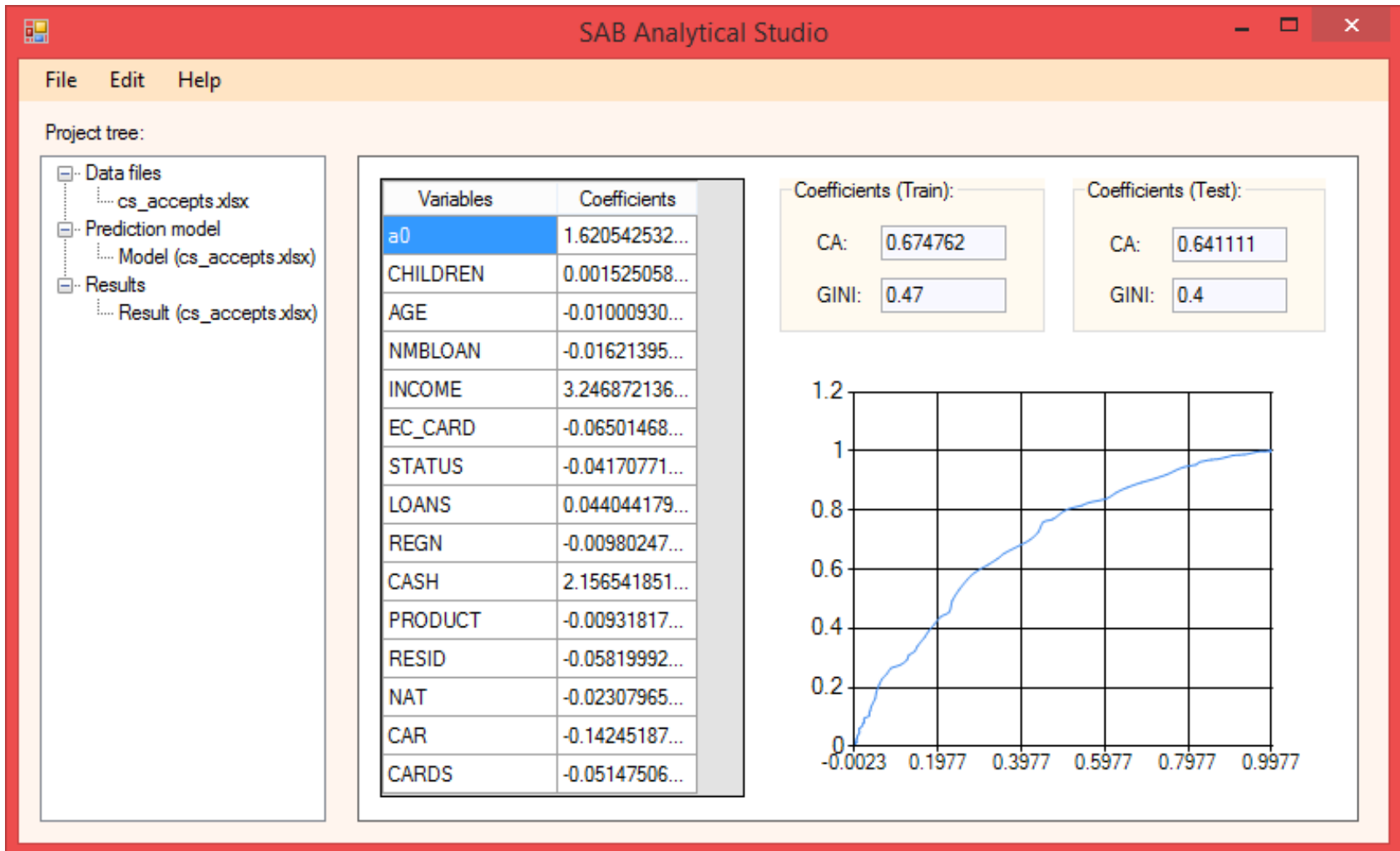
- Для побудови скорингової моделі було використано вибірку даних з німецького банку, що надає кредити фізичним особам.
- Набір даних містить 3000 записів по клієнтах, у яких вже закінчився строк кредитування, та включає в себе інформацію щодо 17 показників (анкетних даних) по кожній особі.

Назва змінної	Опис
TITLE	Стать особи
CHILDREN	Кількість дітей
AGE	Вік
NMBLOAN	Кількість кредитів, наданих цим банком
FINLOAN	Кількість закритих кредитів
INCOME	Дохід особи
EC_CARD	Наявність банківської карти
STATUS	Сімейний стан
LOANS	Кількість відкритих кредитів
REGN	Регіон
CASH	Запитувані грошові кошти
PRODUCT	Ціль кредиту
RESID	Тип місця проживання
NAT	Національність особи
CAR	Наявність транспортного засобу
CARDS	Тип банківської карти
GB	«Хороший» чи «поганий» клієнт

Побудована лінійна імовірнісна модель (з використанням порядкової нумерації категоріальних змінних)



Побудована лінійна імовірнісна модель (з використанням значень коефіцієнту WOE)

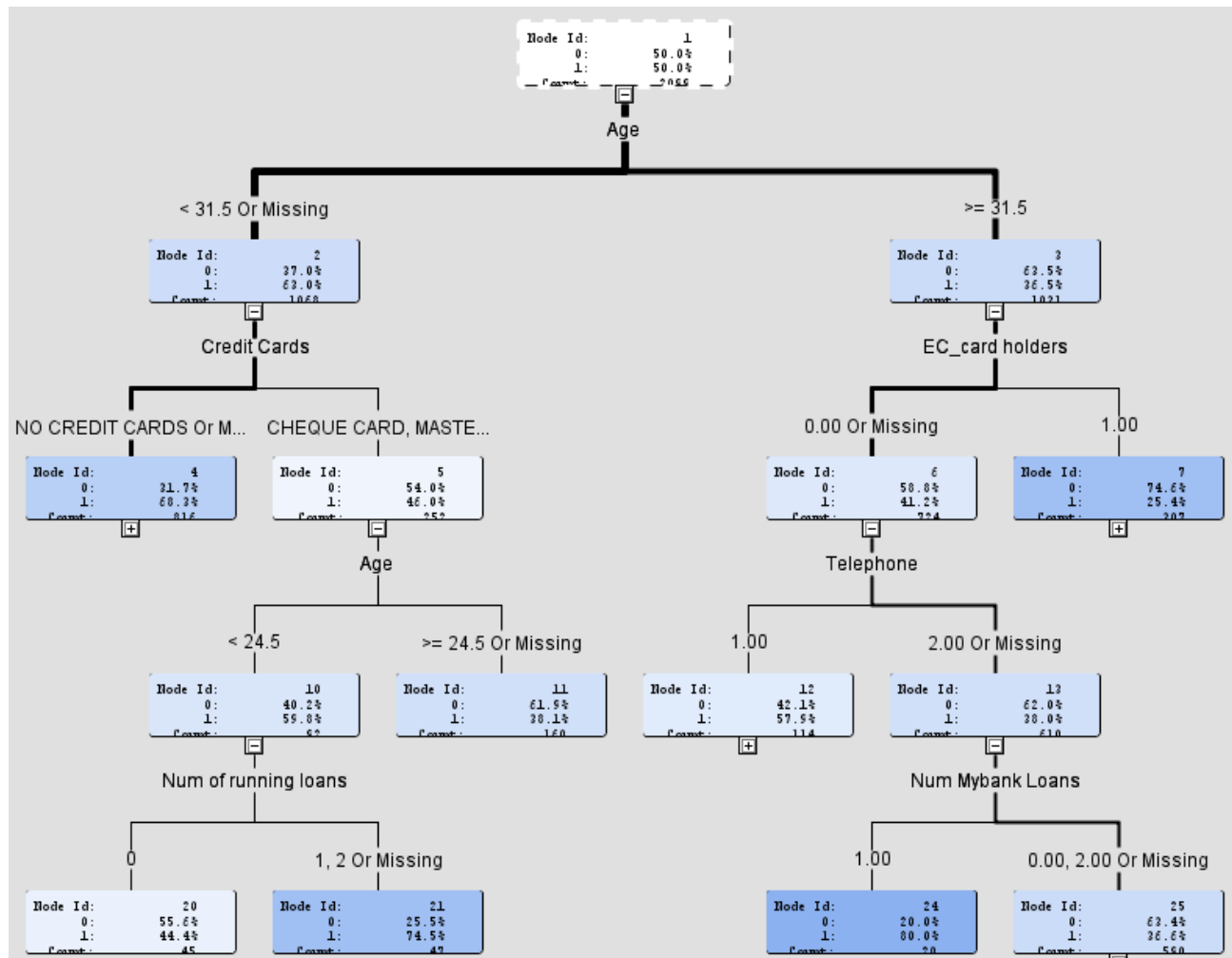


Порівняльна таблиця характеристик побудованих скорингових моделей з використанням WOE та без використання WOE

	Навчальна вибірка (Train)		Тестова вибірка (Test)	
	Загальна точність (CA)	Індекс GINI	Загальна точність (CA)	Індекс GINI
Модель без використання WOE	0.658571	0.45	0.62	0.37
Модель з використанням WOE	0.674762	0.47	0.641111	0.4

- Як можна побачити з таблиці, використання значень коефіцієнту WOE для трансформації категоріальних змінних у числові значно покращує прогнозуючу здатність моделі.

Отримана модель за допомогою методу дерев рішень в системі SAS Enterprise Miner



Отримана скорингова карта на основі логістичної регресії в системі SAS Enterprise Miner

Scorecard

		Scorecard Points
Age	AGE< 23	-10
	23<= AGE< 28	1
	28<= AGE< 31	11
	31<= AGE< 46, _MISSING_	19
	46<= AGE	31
Credit Cards	CHEQUE CARD, MASTERCARD/EUROC, OTHER CREDIT CAR	32
	AMERICAN EXPRESS, NO CREDIT CARDS, VISA MYBANK, VISA OTHERS, _MISSING_, _UNKNOWN_	5
EC_card holders	0.00, _MISSING_, _UNKNOWN_	15
	1.00	6
Income	INCOME< 1000, _MISSING_	18
	1000<= INCOME< 1900	7
	1900<= INCOME< 2500	9
	2500<= INCOME< 3000	12
	3000<= INCOME	15

Аналіз отриманих результатів та порівняння моделей

- Порівняльна таблиця характеристик для скорингових моделей:

	Навчальна вибірка (Train)		Тестова вибірка (Test)	
	Загальна точність (CA)	Індекс GINI	Загальна точність (CA)	Індекс GINI
Лінійна імовірнісна модель (SAB Analytical Studio)	0.674762	0.47	0.641111	0.4
Дерева рішень (SAS Enterprise Miner)	0.686041	0.445	0.681465	0.403
Логістична регресія (SAS Enterprise Miner)	0.665079	0.476	0.622642	0.419

- За критерієм CA найкращою виявилась модель у вигляді дерев рішень.
- За індексом GINI найкращою опинилась логістична регресія.
- Лінійна імовірнісна модель показала середні результати за обома критеріями, що вказує на її хорошу прогнозуючу здатність.

Висновки

- Розглянуто такі методи моделювання ймовірності повернення кредитів: лінійна ймовірнісна модель, дерева рішень, логістична регресія.
- Запропоновано архітектуру системи підтримки прийняття рішень для оцінювання кредитоспроможності фізичних осіб.
- Розроблено програмний продукт SAB Analytical Studio за допомогою мови програмування C# для побудови лінійної ймовірнісної моделі з використанням МНКО для оцінювання параметрів моделі.
- Побудовано математичні моделі у вигляді скорингової карти на основі логістичної регресії, а також дерев рішень із використанням системи SAS Enterprise Miner.
- Виконано порівняльний аналіз якості моделей на основі статистичних критеріїв (CA, індекс GINI).

Перспективи щодо подальших досліджень

- Розробка та реалізація власної комп'ютерної програми, яка б включала в себе побудову логістичної моделі із використанням методу максимальної вірогідності.
- Розробка нових підходів щодо визначення ступеня значущості регресорів.
- Реалізація методів інтелектуального аналізу даних іншого типу, наприклад, нейронні мережі, метод групового врахування аргументів, метод опорних векторів.

Публікації

- Публікація наукової статті у фаховому журналі:
 - “Наукові вісті НТУУ “КПІ””, 2016 р.:
“Методика побудови скорингових карт із використанням платформи SAS”
Бакун С.А., Бідюк П.І.
- Опубліковано тези доповідей на конференціях:
 - 18-та міжнародна науково-технічна конференція SAIT 2016:
“Using SAS Enterprise Miner to build scoring card to evaluation solvency of individuals”
Бакун С.А., Литвинюк А.А., Терентьев О.М.
 - Всеукраїнська науково-практична Інтернет-конференція
“Інформаційно-комунікаційні технології навчання” , 2016р.:
“Using of credit scoring to deciding on the loan”
Бакун С.А., Литвинюк А.А., Терентьев О.М.

Дякую за увагу!
