Тема: «Побудова скорингових моделей за допомогою дискримінантного аналізу»

ВИКОНАВ:
СТУДЕНТ ГРУПИ КА-34
КУДРЯВЦЕВ АНТОН МИХАЙЛОВИЧ

НАУКОВИЙ КЕРІВНИК:
Д.Т.Н. , ПРОФ. БІДЮК ПЕТРО ІВАНОВИЧ

Актуальність проблеми

- Зростаючі темпи розвитку економіки призводять до збільшення попиту на отримання кредитів. Разом з тим збільшується частка неповернень запозичених коштів;
- Вплив умов економічної кризи на зростання частки неповернення кредитів;
- Ручна обробка або автоматизація низького рівня призводить до збільшення затраченого часу на видачу кредиту;
- У зв'язку з цим виникає необхідність побудови сучасних інформаційно-аналітичних систем для аналізу даних позичальників та прогнозування можливості повернення ними кредитів.

Мета роботи:

- Аналіз задач, пов'язаних з оцінюванням кредитного ризику, розробка програмного продукту для оцінювання кредитоспроможності на основі дискримінантного аналізу.

Об'єкт дослідження:

- Вибірка з 10257 підприємств, що отримали кредит. Вибірка являє собою набір змінних, на основі яких проводиться побудова математичної моделі і прогнозування;

Предмет дослідження:

- Скорингові моделей для вирішення задачі прогнозування виплати чи невиплати позичальником кредиту.

Постановка задачі

- Виконати огляд існуючих методів класифікації об'єктів на дві групи, розробити програмний продукт для аналізу даних позичальників кредитів;
- Створити модель для підтримки прийняття рішень за допомогою розробленого програмного продукту на вибірці даних підприємств-клієнтів;
- Також розробити математичну класифікаційну модель за допомогою наявного програмного забезпечення в SPSS;
- Виконати аналіз отриманих результатів, надати рекомендації стосовно видачі кредитів.

Кредитний скоринг

Кредитний скоринг – методика аналізу даних з використанням математичних або статистичних моделей, за допомогою якої на підставі статистичних даних кредитної історії «минулих» клієнтів банк намагається визначити, наскільки велика вірогідність того, що окремий потенційний позичальник поверне кредит в обумовлений строк.

Види скорингу

- Аплікаційний (application scoring)
- Поведінковий (behavioral scoring)
- Колекторський (collection scoring)
- Шахрайства (fraud scoring)

Можливі моделі кредитного скорингу

Основою скорингового аналізуможуть виступати такі моделі:

- Лінійні ймовірнісні моделі;
- Нелінійні Probit і Logit моделі;
- Моделі, побудовані на основі дискримінантного аналізу;
- Нейронні мережі
- Дерева рішень

Дискримінантний аналіз

Розглядається завдання розподілу на два або більше класи досліджуваної вибірки із індивідуумів або об'єктів.

Вирішується задача класифікації клієнтів на дві категорії: «надійний» клієнт і «ненадійний» клієнт, користуючись при цьому даними про величину різниці між деякими з характеристик клієнта.

Отже, необхідно визначити, які з характеристик позичальника найкраще класифікують його до цільових груп.

Дискримінантний аналіз вирішує задачу класифікації шляхом створення так званих дискримінантних функцій $\lambda' x$,

де λ – вектор коефіцієнтів і вагів, присвоєних критеріям x_i.

Модель визначає ці величини, знаходячи якомога більшу можливу різницю між групами.

Критерії якості моделі

- CA (Common Accuracy) – загальна точність моделі, визначається як відношення вірно спрогнозованих випадків до загальної кількості позичальників.
- FPR (False Positive Rate) – частина хибно позитивних спрогнозованих випадків.
- TPR (True Positive Rate) – частина істино позитивних спрогнозованих випадків.
- Помилки першого і другого роду.

Навчальні дані

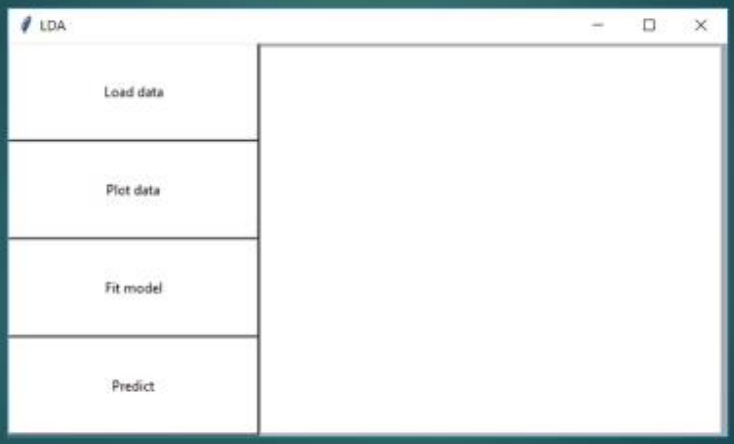
Масив даних являє собою інформацію про 10257 підприємств-клієнтів банку.

Прогнозована змінна - SeriousDlqin2yrs – містить інформацію про факт настання дефолту.

Незалежними змінними обрано:

- MonthlyIncome –загальний прибуток підприємства на місяць
- NumberOfOpenCreditLinesAndLoans - кількість відкритих кредитів (іпотека і/або кредит на автомобіль) і кредитні лінії (наприклад, кредитні карти).
- NumberOfTimes90DaysLate - Кількість прострочень довжиною в 90+ днів.
- NumberRealEstateLoansOrLines - кількість кредитів (іпотека, на покупку нерухомості, кредитні лінії)

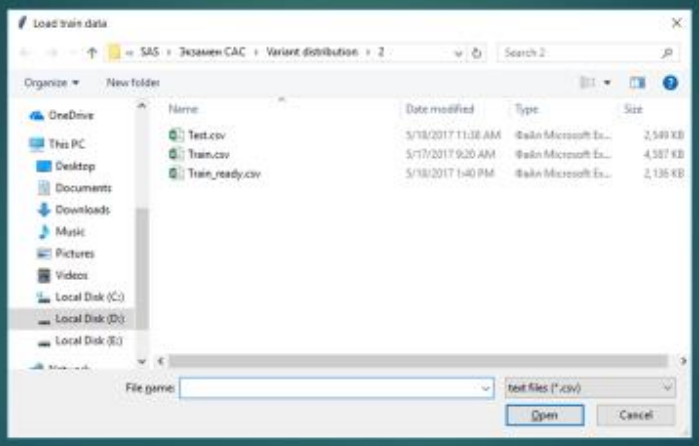
Головний екран програми



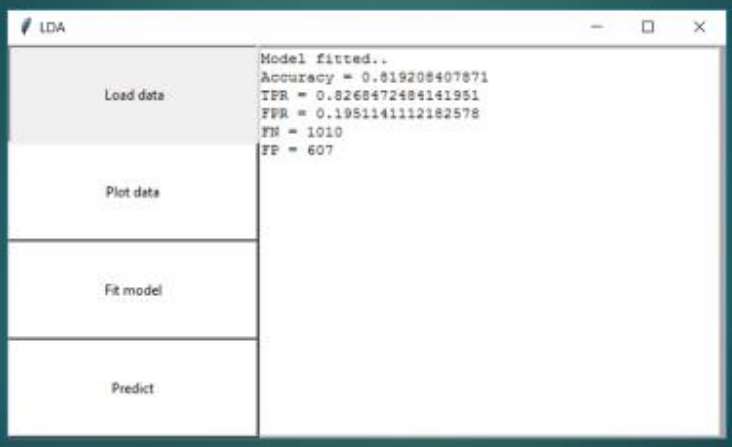
Структура розробленої програми

- Завантаження даних.
- Навчання моделі на основі завантажених даних.
- Використання моделі для прогнозування цільової змінної на тестовому наборі даних.
- Застосування статистичних критеріїв якості результатів.
- Порівняння результатів.

Завантаження даних у програму



Навчання моделі



Результати прогнозування за тестовою вибіркою

	A	B	C	D	E	F	G
1	MonthlyIn	NumberOf	NumberOf	NumberRe	Pred_Seric	Pred_Prob	Pred_Prob
2	2600	4	0	0	'0'	0.978362	0.021638
3	25000	7	0	1	'0'	0.990929	0.009071
4	3500	3	0	1	'0'	0.977734	0.022266
5	3500	8	0	0	'0'	0.977935	0.022065
6	6501	7	0	2	'0'	0.977513	0.022487
7	12454	13	0	2	'0'	0.981114	0.018886
8	1300	6	0	1	'0'	0.97435	0.02565
9	3280	7	0	1	'0'	0.976144	0.023856
10	2500	4	0	0	'0'	0.978266	0.021734
11	2416	9	0	1	'0'	0.974474	0.025526
12	2500	17	0	0	'0'	0.973655	0.026345
13	2230	7	0	0	'0'	0.977005	0.022995
14	11000	9	0	1	'0'	0.982541	0.017459
15	5700	14	0	1	'0'	0.97622	0.02378
16	9250	4	0	0	'0'	0.98389	0.01611
17	3133	2	0	0	'0'	0.979485	0.020515
18	6900	21	1	1	'1'	0.45227	0.54773
19	2231	2	0	0	'0'	0.978649	0.021351

Побудова моделі в SPSS

На даному слайді показано статистики спостережень, розподілених по групах

Статистика группы

SeriousDlqin2yrs		Среднее	Среднекв. отклонение	N валидных (по списку)	
				Невзвешенные	Взвешенные
0	MonthlyIncome	3343.66	2250.056	3111	3111.000
	NumberOfOpenCreditLinesAndLoans	4.10	2.748	3111	3111.000
	NumberOfTimes90DaysLate	.12	.467	3111	3111.000
	NumberRealEstateLoansOrLines	.03	.194	3111	3111.000
1	MonthlyIncome	5489.36	3915.585	5833	5833.000
	NumberOfOpenCreditLinesAndLoans	8.28	5.218	5833	5833.000
	NumberOfTimes90DaysLate	.58	.956	5833	5833.000
	NumberRealEstateLoansOrLines	.95	1.096	5833	5833.000
Всего	MonthlyIncome	4743.02	3578.193	8944	8944.000
	NumberOfOpenCreditLinesAndLoans	6.82	4.934	8944	8944.000
	NumberOfTimes90DaysLate	.42	.849	8944	8944.000
	NumberRealEstateLoansOrLines	.63	.995	8944	8944.000

Коефіцієнти отриманої канонічної функції дискримінації

Коэффициенты канонической дискриминантной функции

	Функция
	1
MonthlyIncome	.000
NumberOfOpenCreditLinesAndLoans	.110
NumberOfTimes90DaysLate	.754
NumberRealEstateLoansOrLines	.597
(Константа)	-1.613

Результати класифікації

Результаты классификации[a]

			Предсказанная принадлежность к группе		
		SeriousDlqin2yrs	0	1	Всего
Исходный	Количество	0	2504	607	3111
		1	1010	4823	5833
	%	0	80.5	19.5	100.0
		1	17.3	82.7	100.0

a. 81.9% исходных сгруппированных наблюдений классифицированы правильно.

Висновки

- Виконано аналіз задачі оцінювання кредитного ризику на основі статистичних характеристик позичальників. Встановлена необхідність побудови сучасних інформаційно-аналітичних систем для автоматизованого розв'язання поставленої задачі.
- Розроблено програмний продукт у системі Python для прийняття рішень стосовно видачі кредиту за методом скорингової моделі;
- Виконано аналіз результатів застосування розробленого програмного продукту, а також аналіз отриманих результатів побудови скорингової моделі в системі SPSS.
- Показано, що розроблений програмний продукт забезпечує необхідну якість прогнозування дефолту позичальників кредитів.

РЕКОМЕНДАЦІЇ ДО ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

В рамках подальших досліджень рекомендовано виконати подальшу адаптацію моделі до даних з метою досягнення більшої точності, а також реалізувати моделі на основі інших методів (логіт, пробіт моделі, дерева рішень, байєсівські мережі) і застосувати комбінування оцінок прогнозів.

Дякую за увагу!

Кожна група має пару векторів (μ_1, Σ_1) і (μ_2, Σ_2), присвоєні їм, які відображають математичне сподівання групи і коваріацію, відповідно.

Нехай p_i - ймовірність того, що окремий заявник відноситься до групи i,

c_{ij} - витрати, понесені у зв'язку з неправильною класифікацією, коли заявник в групі i повинен бути переданий групі j.

Тоді заявник з присвоєним йому вектором характеристик x, класифікується до групи G_1, якщо

$$\lambda' x \geq \alpha + ln\left(\frac{c_{21}p_2}{c_{12}p_1}\right),$$

де $\lambda = \Sigma^{-1}(\mu_1 - \mu_2)$,

$$\alpha = \frac{\lambda(\mu_1 + \mu_2)}{2}.$$

В інших випадках заявник відноситься до групи G_2.