Аналіз кредитоспроможності клієнтів за допомогою нелінійних ймовірнісних моделей

Виконав: студент групи КА-34

Савенков Юрій Ростиславович

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

Актуальність роботи

- Разом із зростаючими темпами розвитку економіки, що призводить до збільшення попиту на кредитування, збільшується частка неповернень запозичених у банку коштів;
- Системи підтримки прийняття рішень низького рівня та ручна обробка даних призводить до затримок, збільшення затраченого часу та меншої вірогідності повернення кредиту;
- Вплив нестабільних економічних умов на невиплату кредиту;
- Зважаючи на ці фактори, виникає необхідність побудови сучасних систем для аналізу даних позичальників та прогнозування можливості повернення ними запозичених коштів.

Об’єкт, предмет, мета

- Об’єкт дослідження: Вибірка з 10437 підприємств, що раніше отримали кредит. Ця вибірка включає в себе параметри, на основі яких проводиться побудова математичної моделі та прогнозування;
- Мета дослідження: Аналіз та оцінювання задач кредитного ризику. Розробка програмного продукту на основі моделі логістичної регресії для оцінювання кредитоспроможності підприємств - майбутніх позичальників;
- Предмет дослідження: Прогнозування за допомогою скорингової моделі ймовірності виплати чи невиплати позичальником кредиту. Реалізація програмного продукту виконана в середовищі Python 3.5. Порівняння отриманих результатів було виконане з побудованою в продукті IBM SPSS моделлю логістичної регресії.

Постановка задачі

- Провести огляд існуючих методів аналізу кредитоспроможності підприємства та розробити програмний продукт для аналізу даних позичальників;
- На основі даних підприємств-клієнтів банку створити модель для підтримки прийняття рішень;
- Побудувати математичну модель на основі логістичної регресії в продукті IBM SPSS для проведення порівняльного аналізу;
- Проаналізувати отримані результати та надати рекомендації стосовно видачі кредиту.

Кредитний скоринг

Скоринг визначається як сукупність статистичних методів, за допомогою яких на підставі статистичних даних кредитної історії «минулих» клієнтів банк намагається визначити, наскільки велика вірогідність того, що окремий потенційний позичальник поверне кредит у встановлений строк.

Логістична регресія

- Бінарна логістична регресія застосовується в тому випадку, коли вихідна змінна може приймати тільки два значення;
- Найчастіше в задачах кредитування використовується саме бінарна логістична регресія, оскільки значення вихідної змінної (повернення кредиту) може набувати двох значень: повернення кредиту (0 або true = Non_default), або неповернення кредиту (1 або False = Default);
- У випадку застосування бінарної логістичної регресії ймовірність p входження клієнта у стан дефолту D визначається як $p = \frac{\exp(\beta X)}{1+\exp(\beta X)}$.

Методи оцінки якості моделі

- CA (Common Accuracy) – загальна точність моделі. Визначається як співвідношення кількості вірно спрогнозованих випадків до загального числа позичальників;
- TPR (True Positive Rate) – частка правильно класифікованих позитивних клієнтів;
- FPR (False Positive Rate) – частка неправильно класифікованих негативних клієнтів;
- FN (False Negatives) — позитивні клієнти, класифіковані як негативні (похибка I роду);
- FP (False Positives) — негативні клієнти, класифіковані як позитивні (похибка II роду).

Результат класифікації моделі	**Фактична приналежність**	
	Позитивний	Негативний
Позитивний	TP (істинно позитивні)	FP (хибно позитивні)
Негативний	FN (хибно негативні)	TN (істинно негативні)

Вибірка даних

Масив даних являє собою інформацію про 10437 підприємств-клієнтів банку, що раніше отримали кредит.

Прогнозована змінна - SeriousDlqin2yrs – це факт настання дефолту.

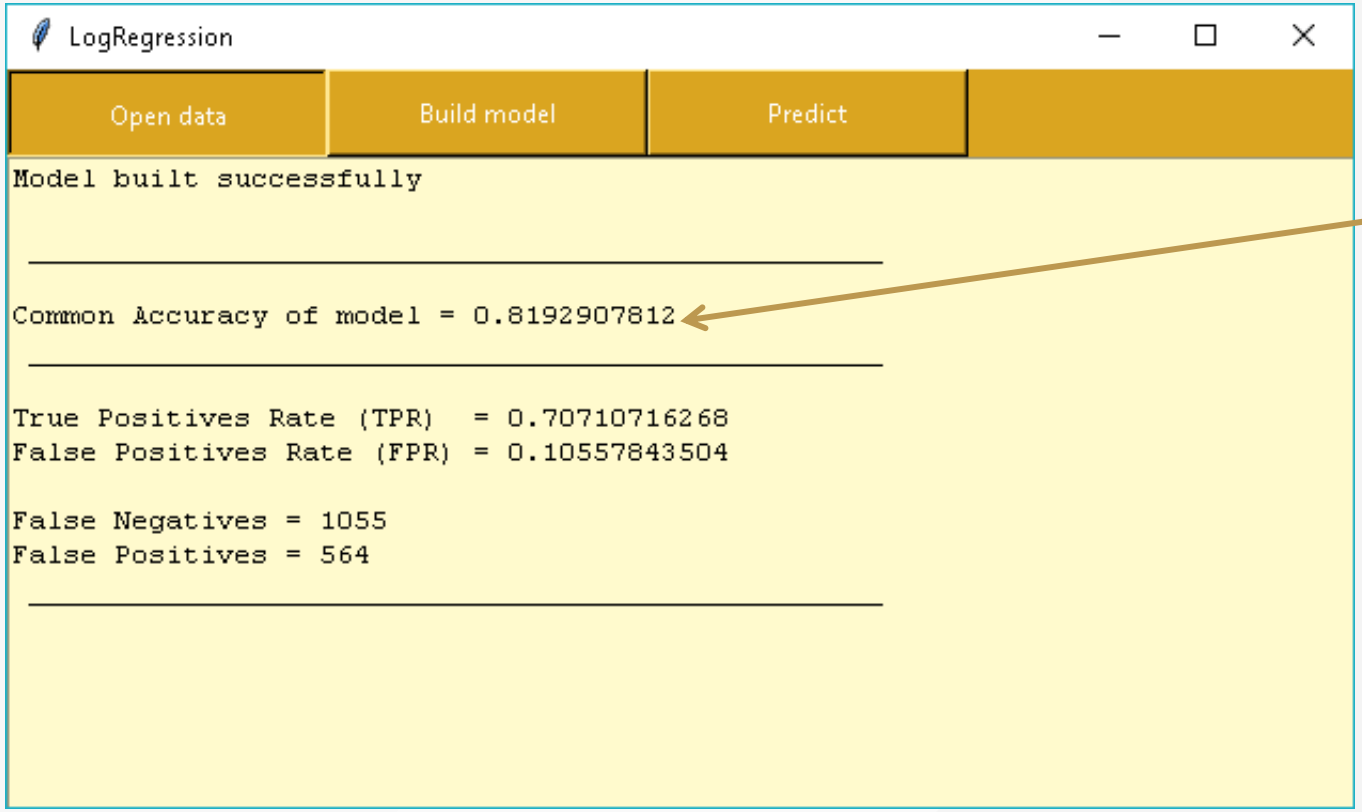
Незалежні змінні:

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

Функціонал програми

- Можливість завантаження власних даних;
- Навчання моделі на основі завантажених даних;
- Використання моделі логістичної регресії для прогнозування цільової змінної на тестовому наборі даних;
- Застосування статистичних критеріїв для оцінки якості результатів;
- Порівняння отриманих результатів.

Результати навчання моделі



Загальна точність моделі

Результати прогнозування

Спрогнозована змінна

MonthlyIr	NumberO	NumberO	NumberR	Pred_Seri	Pred_Prob	Pred_Prob
6140	6	0	0	'0'	0.607709	0.392291
6579	15	0	2	'1'	0.208994	0.791006
5500	11	0	1	'1'	0.011195	0.988805
6480	12	1	1	'1'	0.134867	0.865133
4466	11	0	1	'1'	0.054163	0.945837
15000	18	0	2	'1'	0.380177	0.619823
7823	3	0	0	'1'	0.157144	0.842856
8500	17	0	4	'1'	0.342862	0.657138
10300	7	0	0	'1'	0.131513	0.868487
1300	6	0	1	'1'	0.146543	0.853457
2253	3	0	0	'0'	0.555733	0.444267
8200	3	0	1	'0'	0.572963	0.427037
1641	9	0	2	'1'	0.180307	0.819693
2343	5	0	0	'1'	0.256385	0.743615
4570	24	0	2	'1'	0.143868	0.856132
2500	7	0	1	'1'	0.030317	0.969683
9500	10	1	1	'1'	0.174242	0.825758
10500	19	0	2	'1'	0.002491	0.997509
1300	2	0	0	'1'	0.271021	0.728979
3000	2	0	0	'1'	0.274302	0.725698
7150	3	0	2	'1'	0.31587	0.68413
5200	4	0	1	'1'	0.150011	0.849989
9116	6	0	2	'0'	0.717973	0.282027
5200	14	0	3	'1'	0.072018	0.927982

Результати класифікації в програмному продукті IBM SPSS

Classification Table[a]

Observed			Predicted		
			SeriousDlqin2yrs		Percentage Correct
			0	1	
Step 1	SeriousDlqin2yrs	0	2547	564	81.9
		1	1055	4778	81.9
	Overall Percentage				81.9

Отже, можна побачити, що отримані результати за допомогою розробленого програмного продукту із результатами, отриманими в системі SPSS мають однакову точність, тому розроблений програмний продукт надає можливість будувати якісну скорингову модель на основі логістичної регресії.

Висновки

- Було розроблено програмний продукт для аналізу та оцінювання кредитоспроможності клієнтів;
- Проведено аналіз задачі оцінки кредитного ризику на основі статистичних характеристик позичальників. Виявлена необхідність побудови сучасних інформаційно-аналітичних систем для автоматизованого розв’язання поставленої задачі;
- Виконано аналіз результатів, отриманих за допомогою розробленого програмного продукту, а також проведений порівняльний аналіз отриманих результатів в системі IBM SPSS;
- Встановлено, що розроблений програмний продукт забезпечує необхідну якість прогнозування дефолту підприємств-позичальників.

Рекомендації до подальших досліджень

Подальшим дослідженням для цієї роботи могли б стати подальша адаптація моделі для досягнення більшої точності та впровадження в програмному продукті інших методів прогнозування (дерева рішень, байєсівські мережі, дискримінантний аналіз) і застосування комбінування оцінок.

Дякую за увагу!

Логістична регресія

У випадку застосування бінарної логістичної регресії ймовірність p входження клієнта у стан дефолту D визначається як

$$p = \frac{\exp(\beta X)}{1+\exp(\beta X)}.$$

Очевидно, що під час застосування множинної регресії для визначення ймовірності повернення кредиту позичальника може виявитись, що ймовірність дефолту буде знаходитися поза межами інтервалу [0; 1].

Якщо вибирається логістична функція розподілу для подання ймовірності відклику p, то така трансформація веде до переходу до логіт-моделі. В цьому випадку буде

$$p = \Phi(b'x) = \int_{-\infty}^{b'x} \varphi(z)dz = \frac{1}{1+e^{-b'x}}$$

або в іншому вигляді

$$p = \frac{e^{b_1x_1+\cdots+b_nx_n}}{1+e^{b_1x_1+\cdots+b_nx_n}}.$$