

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»
НАВЧАЛЬНО-НАУКОВИЙ КОМПЛЕКС «ІНСТИТУТ
ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ»
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Оцінювання кредитоспроможності клієнтів банку

Виконав: студент групи КА-23 Моругий І.А.

Науковий керівник: проф. Бідюк П.І.

Київ 2016

Об'єкт, предмет і мета дослідження

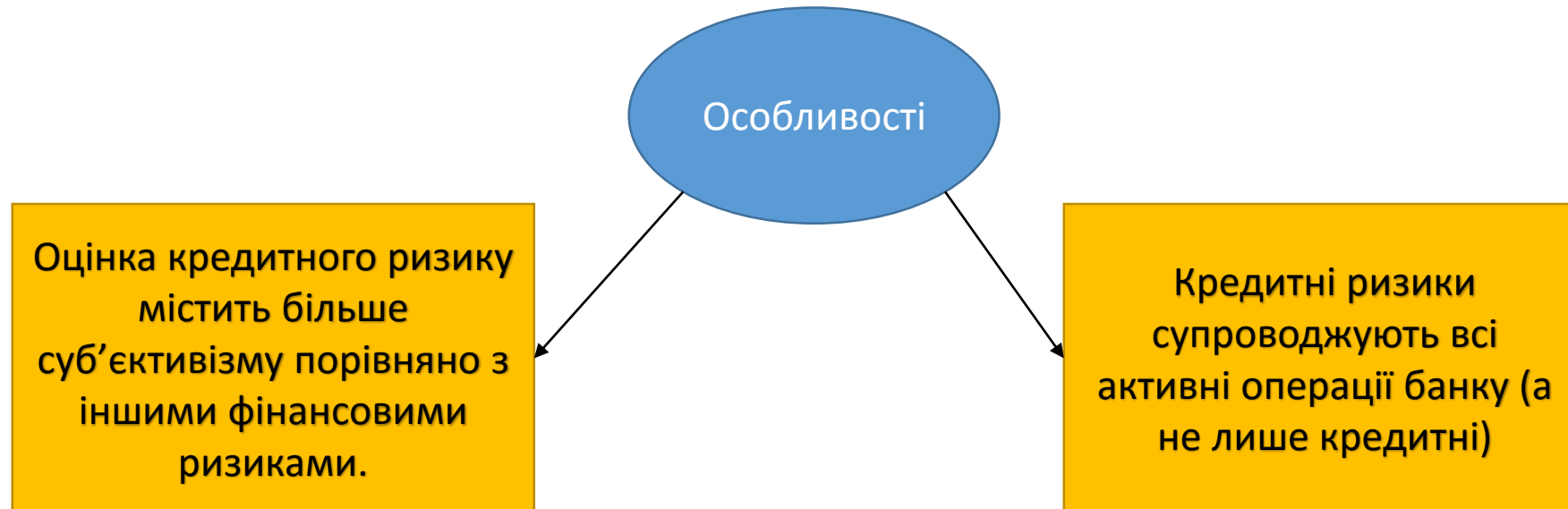
- Об'єкт дослідження: клієнти банку – позичальники кредитів, представлені статистичними даними стосовно їх характеристик.
- Предмет дослідження: Математичні моделі і методи оцінювання і прогнозування кредитного ризику.
- Мета дослідження: Побудова математичних моделей кредитного ризику методами інтелектуального аналізу даних, розробити необхідне програмне забезпечення.

Постановка задачі

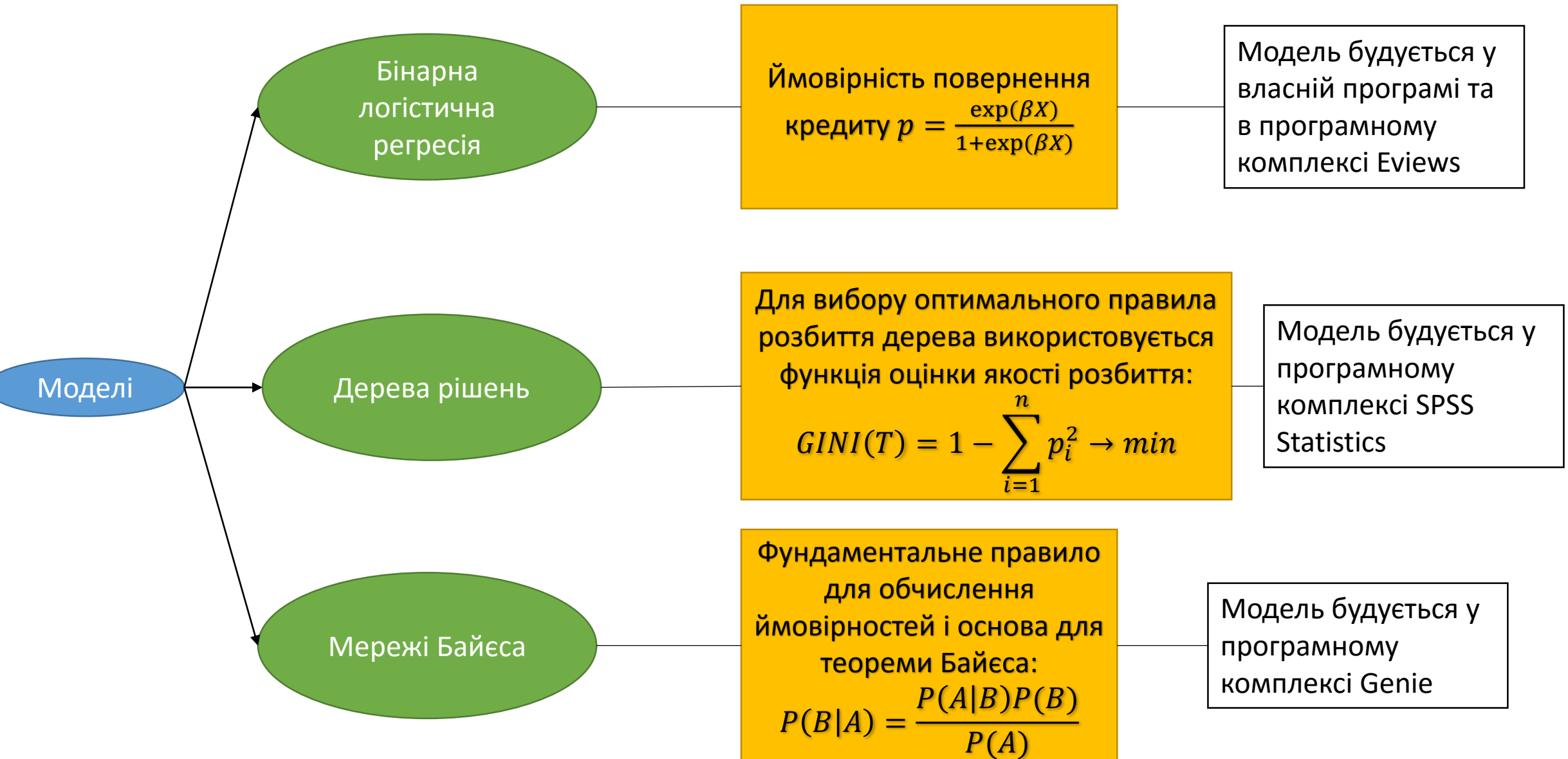
1. Обґрунтувати поняття кредитного ризику та вказати особливості його виникнення.
2. Побудувати вибрані математичні моделі опису кредитного ризику: бінарна логістична регресія, дерева рішень, мережі Байєса.
3. Розробити програмний продукт, призначений для побудови моделей оцінювання кредитного ризику.
4. Проаналізувати отримані моделі стосовно якості прогнозування.

Кредитний ризик та його особливості

- Кредитний ризик – це ризик банку який кредитує, що полягає в тому, що в період, протягом якого позика залишатиметься непогашеною, позичальник буде нездатний виконати умови кредитної угоди, що лежить в основі позички.



Математичні моделі опису кредитних ризиків використані в дипломній роботі



Об'єкт дослідження

- Для побудови математичних моделей використовується вибірка з даних 15000 клієнтів банку (фізичних осіб).

Назва	Значення	Тип
Стать	0-жінка; 1-чоловік	Регресор
Вік	У повних роках	Регресор
Сума кредиту	В гривнях	Регресор
Строк кредиту	У днях	Регресор
Сімейний стан	0 - не одружений; 1 - громадянський шлюб; 2 - одружений; 3 - розлучений.	Регресор
Кількість дітей		Регресор
Кількість працівників у компанії		Регресор
Дохід	В гривнях	Регресор
Витрати	В гривнях	Регресор
Результат	0 - не повернув кредит; 1 – повернув кредит.	Залежна змінна

Логіт-моделювання: побудова моделі

Equation: UNTITLED Workfile: LOGIT::Untitled\

View Proc Object Print Name Freeze Estimate Forecast Stats Resids

Dependent Variable: RESULT
Method: ML - Binary Logit (Quadratic hill climbing)
Date: 05/06/16 Time: 22:26
Sample: 1 15000
Included observations: 15000
Convergence achieved after 10 iterations
Covariance matrix computed using second derivatives

Variable	Coefficient	Std. Error	z-Statistic	Prob.
GENDER	0.344130	0.075805	4.539669	0.0000
AGE	-0.011070	0.003908	-2.832976	0.0046
CREDITSUM	2.81E-06	1.08E-06	2.602916	0.0092
TERMCREDITINGINDAY	-1.21E-08	1.55E-08	-0.782811	0.4337
MATIRIALSTATUS	0.053798	0.039973	1.345852	0.1784
CHILDREN	0.083359	0.057502	1.449659	0.1472
EMPLOYEESINCOMP	-0.104553	0.027120	-3.855164	0.0001
INCOMECUSTOMER	-1.32E-06	1.42E-06	-0.933947	0.3503
COSTSCUSTOMER	2.19E-05	2.99E-05	0.732781	0.4637
C	3.269787	0.197731	16.53650	0.0000

McFadden R-squared	0.012663	Mean dependent var	0.950000
S.D. dependent var	0.217952	S.E. of regression	0.217350
Akaike info criterion	0.393336	Sum squared resid	708.1427
Schwarz criterion	0.398414	Log likelihood	-2940.023
Hannan-Quinn criter.	0.395021	Deviance	5880.046
Restr. deviance	5955.457	Restr. log likelihood	-2977.729
LR statistic	75.41174	Avg. log likelihood	-0.196002
Prob(LR statistic)	0.000000		

Obs with Dep=0	750	Total obs	15000
Obs with Dep=1	14250		

Логіт-моделювання

Оцінка прогнозу

Успіх, якщо ймовірність більша ніж

Коеф. навч. вибірки

Побудувати

Параметри

Логарифм Функції Правдоподібності (ЛФП)

-2 ЛФП

Тест відношення правдоподібності (LR statistic)

Дод. параметри

Додаткові параметри

	Змінні	Коефіцієнти	Стандартне відхилення
▶	intercept	3,26978750669803	0,197736490944884
	Input 0	0,344129483483322	0,0758069072035618
	Input 1	-0,0110699523960559	0,00390753873005388
	Input 2	2,81083889255362E-06	1,07987885008455E-06
	Input 3	-1,21389841991387E-08	1,55069080783088E-08
	Input 4	0,0537976453679624	0,0399729455150215
	Input 5	0,0833590046540103	0,0575024732388248
	Input 6	-0,104553203278341	0,0271202314194179
	Input 7	-1,32170392892244E-06	1,41518106065807E-06
	Input 8	2,19242681319567E-05	2,99192560291706E-05
*			

Логіт-моделювання: загальна точність

Фактично\Модель	Дефолт (0)	Повернення кредиту (1)	Відсоток точності
Поріг відсікання = 0.95			
Дефолт (0)	467	283	62.26%
Повернення кредиту (1)	6486	7764	54.48%
Загальна точність моделі			54.87%
Поріг відсікання = 0.9			
Дефолт (0)	1	749	0.13%
Повернення кредиту (1)	8	14242	99.94%
Загальна точність моделі			94.95%
Поріг відсікання = 0.85			
Дефолт (0)	0	750	0%
Повернення кредиту (1)	2	14248	99.98%
Загальна точність моделі			94.99%

$$CA = \frac{CorrectForecast}{N}$$

CorrectForecast - кількість правильно спрогнозованих випадків.
N – загальна кількість випадків.

Логіт-моделювання: тест Хосмера-Лемешова

Equation: EQ01 Workfile: LOGIT::Untitled\								
View	Proc	Object	Print	Name	Freeze	Estimate	Forecast	Stats
Resids								
Goodness-of-Fit Evaluation for Binary Specification								
Andrews and Hosmer-Lemeshow Tests								
Equation: EQ01								
Date: 05/08/16 Time: 02:01								
Grouping based upon predicted risk (randomize ties)								
Quantile of Risk		Dep=0		Dep=1		Total	Obs	H-L
Low	High	Actual	Expect	Actual	Expect			
1	0.7590	0.9292	133	118.185	1367	1381.82	1500	2.01609
2	0.9292	0.9369	110	99.9970	1390	1400.00	1500	1.07211
3	0.9370	0.9427	78	90.1380	1422	1409.86	1500	1.73901
4	0.9427	0.9474	84	82.3453	1416	1417.65	1500	0.03518
5	0.9474	0.9514	84	75.7971	1416	1424.20	1500	0.93499
6	0.9514	0.9550	60	70.2138	1440	1429.79	1500	1.55872
7	0.9550	0.9588	50	64.6451	1450	1435.35	1500	3.46723
8	0.9588	0.9632	54	58.5789	1446	1441.42	1500	0.37246
9	0.9632	0.9685	53	51.2960	1447	1448.70	1500	0.05861
10	0.9685	0.9998	44	38.8044	1456	1461.20	1500	0.71413
Total			750	750.000	14250	14250.0	15000	11.9685
H-L Statistic			11.9685	Prob. Chi-Sq(8)		0.1526		
Andrews Statistic			33.9162	Prob. Chi-Sq(10)		0.0002		

Суть тесту полягає у тому, що всі клієнти розбиваються на 10 груп (в кожній групі 10% від загальної кількості клієнтів) – “квантилів ризику”. У кожному квантилі порівнюються справжні та очікувані розподіли частот кредитоспроможних та дефолтних клієнтів. В результаті підраховується статистика Хосмера-Лемешова. У даному випадку отримано значення 11.9685 при рівні значимості 0.1526. Оптимальними вважаються рівні значимості не менше 0.5-0.6.

$$H = \sum_{g=1}^G \frac{(O_{1g} - E_{1g})^2}{E_{1g}} + \frac{(O_{0g} - E_{0g})^2}{E_{0g}}$$

Логіт-моделювання: ROC-аналіз

ROC-аналіз

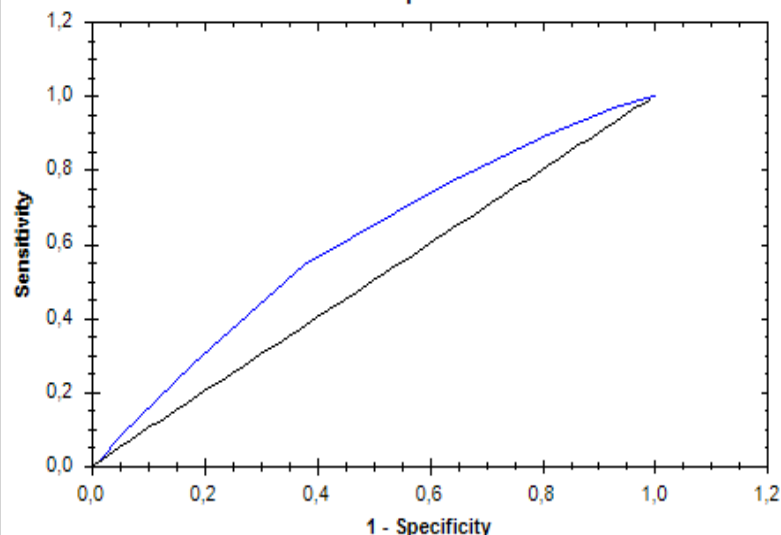
Поріг (від)

Поріг (до)

Крок

Провести

ROC-крива



AUC
(Area Under Curve)

GINI

Пошук оптимального порогу

Результат для выбранного порогу

Поріг

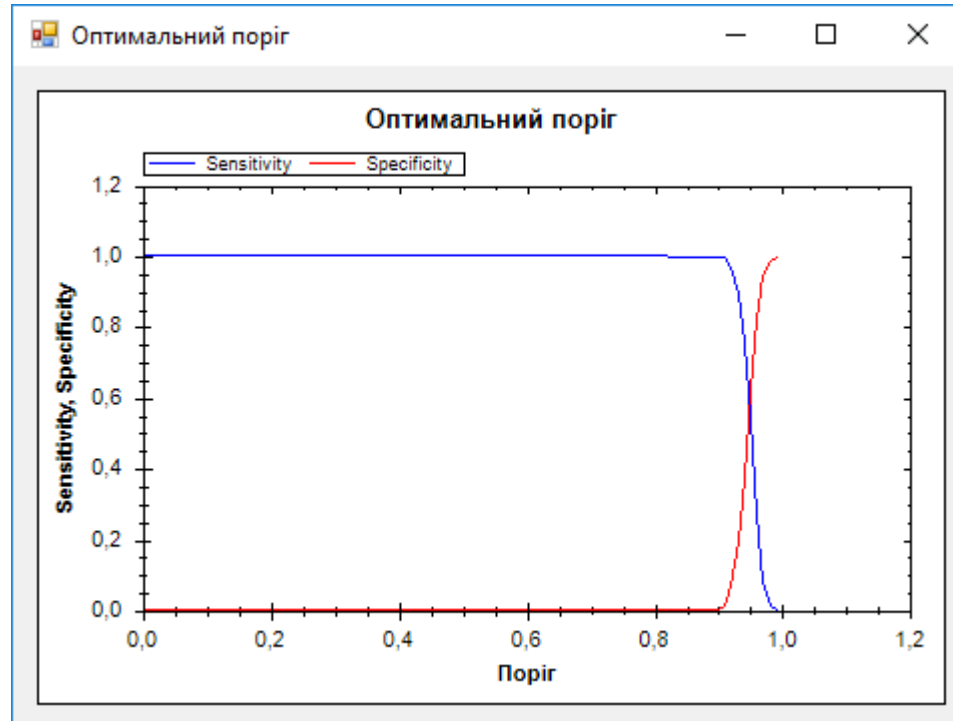
Відобразити

Фактично\Модель	Не поверне (0)	Поверне (1)
Не поверне (0)	467	283
Поверне (1)	6486	7764
Sensitivity:	0,544842105263...	
Specificity:	0,622666666666...	

Інтервал AUC	Інтервал GINI	Якість моделі
0.9 – 1.0	0.8 – 1.0	Відмінна
0.8 – 0.9	0.6 – 0.8	Дуже висока
0.7 – 0.8	0.4 – 0.6	Прийнятна
0.6 – 0.7	0.2 – 0.4	Середня
0.5 – 0.6	0 – 0.2	Незадовільна

Загальний висновок стосовно якості:
Середня.

Логіт-моделювання: оптимальний поріг



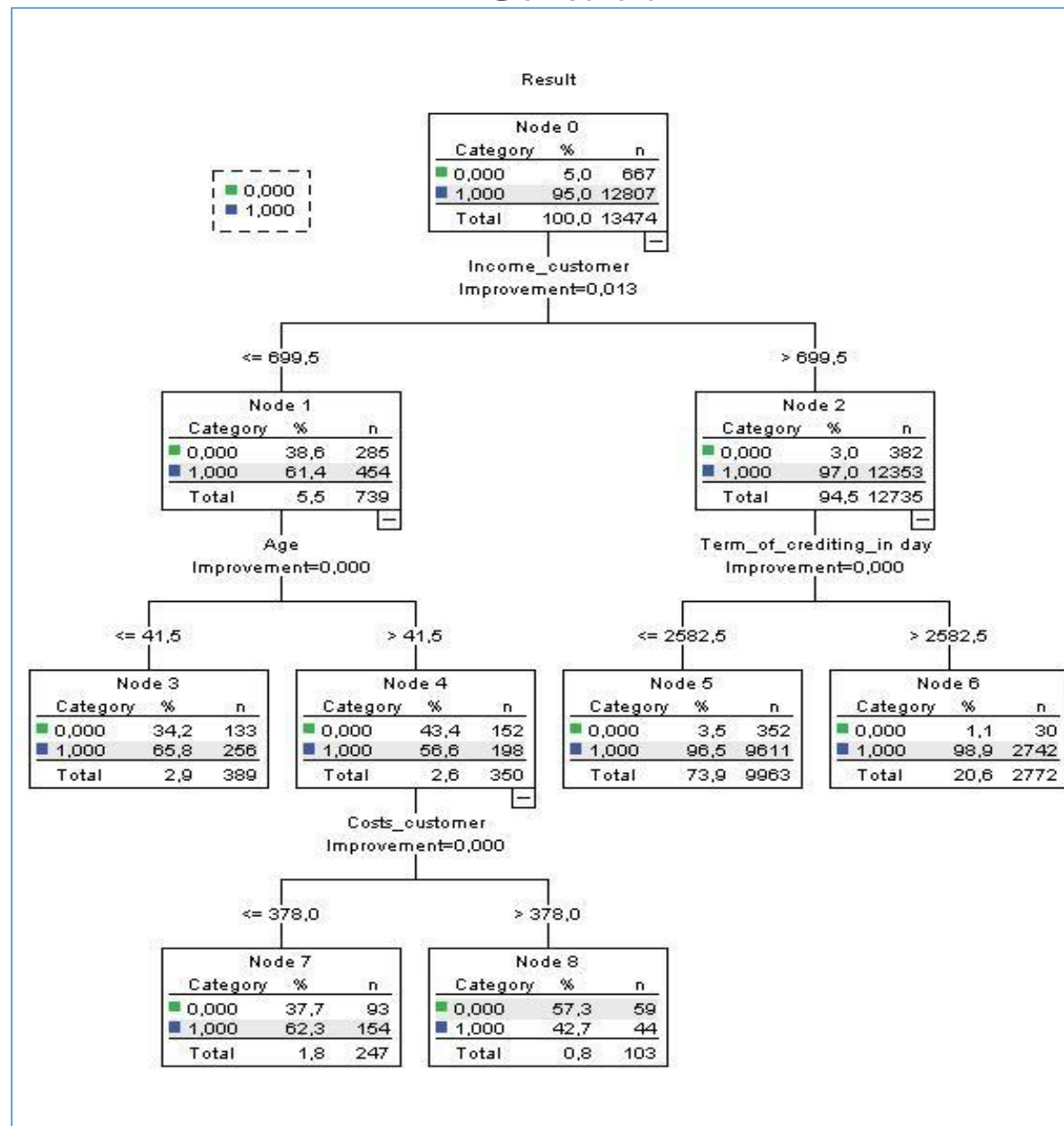
В якості критерію вибору оптимального порогу було обрано досягнення балансу між специфічністю та чутливістю моделі:

$$Cutt_off_o = \min_k |Se_k - Sp_k|$$

Логіт-моделювання: висновок

- Слід відмітити, що вибір кращої моделі залежить від пріоритетів дослідника. Варіюючи значення порогу, можна знайти бажані значення чутливості та специфічності. Якщо більш важливо, щоб модель краще розпізнавала кредитоспроможних клієнтів необхідно збільшувати значення чутливості. Для кращого розпізнавання дефолтників необхідно збільшувати значення специфічності. На практиці неможливо отримати абсолютно ідеальну модель. Чим більша чутливість тим менша специфічність і навпаки.

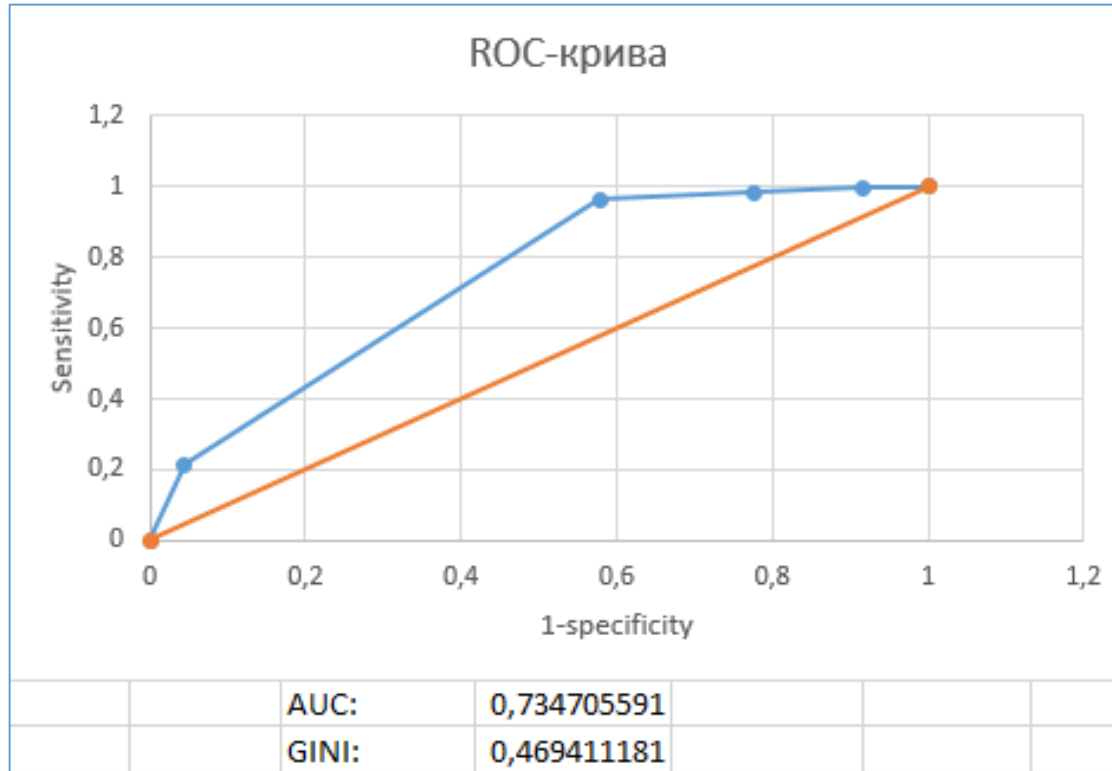
Дерева рішень: побудова моделі алгоритмом CART



Дерева рішень: загальна точність моделі

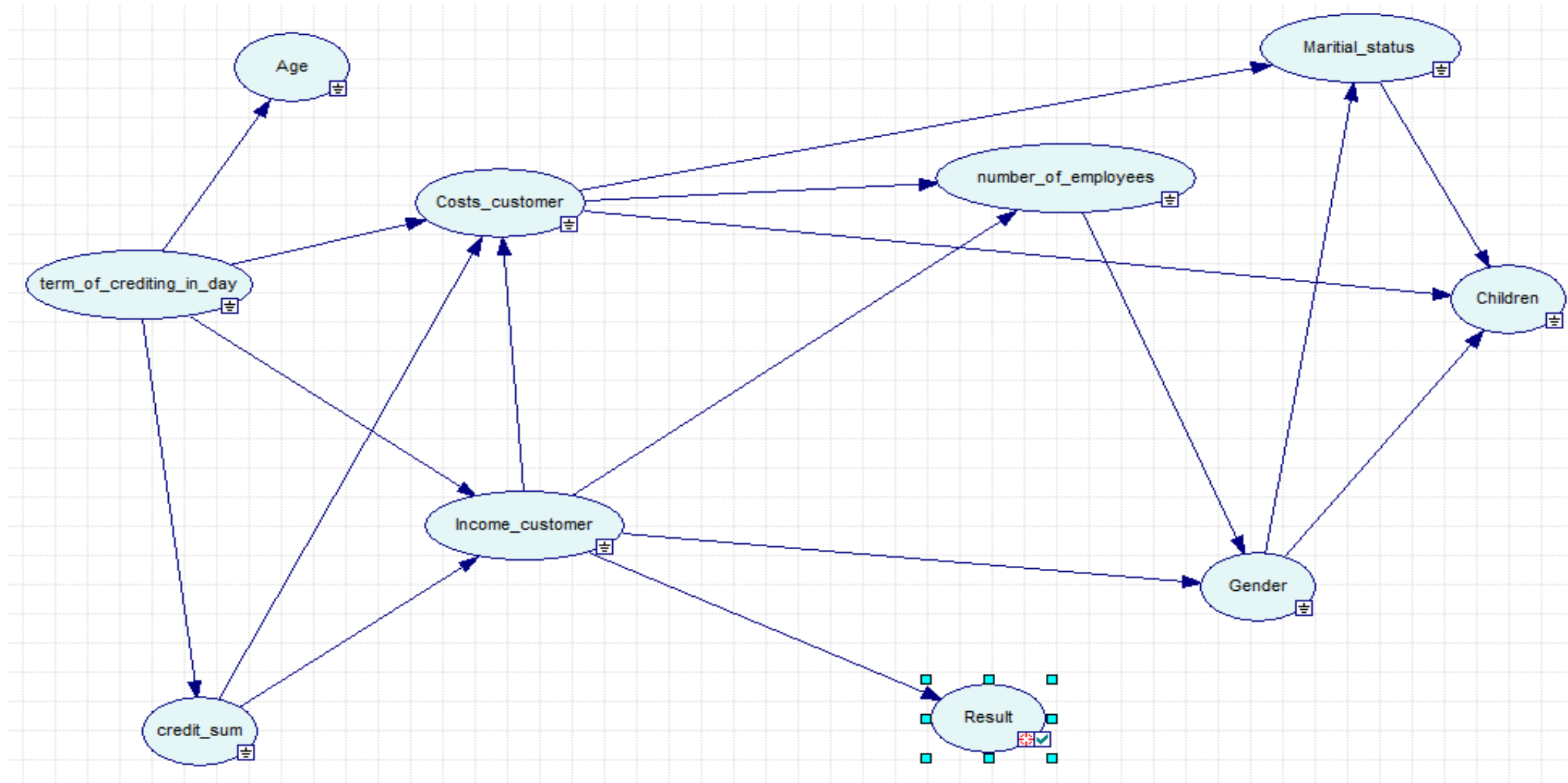
Фактично\Модель	Дефолт (0)	Повернення кредиту (1)	Відсоток точності
Поріг відсікання = 0.95			
Дефолт (0)	316	434	42.13%
Повернення кредиту (1)	508	13742	96.43%
Загальна точність моделі			93.72%
Поріг відсікання = 0.65			
Дефолт (0)	167	583	22.26%
Повернення кредиту (1)	223	14027	98.43%
Загальна точність моделі			94.63%
Поріг відсікання = 0.6			
Дефолт (0)	62	688	8.27%
Повернення кредиту (1)	50	14200	99.64%
Загальна точність моделі			95.08%

Дерева рішень: ROC-аналіз для алгоритму CART



Загальний висновок стосовно якості:
Прийнятна

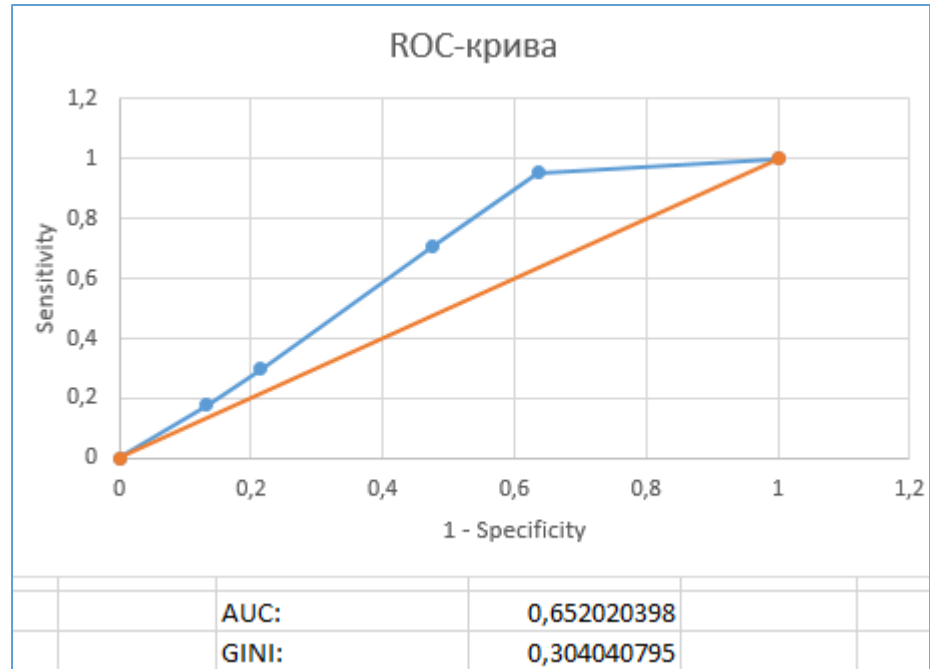
Мережа Байєса: побудова моделі



Мережа Байєса: загальна точність

Фактично\Модель	Дефолт (0)	Повернення кредиту (1)	Відсоток точності
Поріг відсікання = 0.97			
Дефолт (0)	589	161	78.53%
Повернення кредиту (1)	9989	4261	29.9%
Загальна точність моделі			32.33%
Поріг відсікання = 0.95			
Дефолт (0)	273	477	36.4%
Повернення кредиту (1)	657	13593	95.38%
Загальна точність моделі			92.44%
Поріг відсікання = 0.85			
Дефолт (0)	0	750	0%
Повернення кредиту (1)	0	14250	100%
Загальна точність моделі			95%

Мережа Байєса: ROC-аналіз



Загальний висновок стосовно якості:
Середня

Наукова новизна дослідження

- Побудовані нові високоадекватні моделі для оцінювання кредитного ризику фізичних осіб – позичальників кредитів банку у формі логістичної регресії, дерева рішень та мережі Байєса.
- Створено новий програмний продукт на мові C#, Visual Studio, для оцінювання кредитоспроможності позичальників кредитів за методом логістичної регресії.

Висновок

- Під час виконання дипломної роботи побудовані моделі оцінки кредитного ризику фізичних осіб – позичальників кредитів банку. Ці моделі включають в себе логістичну регресію, дерева рішень, мережі Байєса. За результатами кожного з методів було показано, що дерево рішень побудоване за алгоритмом CART показало найкращий результат.
- Щодо варіантів покращення отриманих результатів, то можна запропонувати інший варіант формування вибірки, в якій доля дефолтників буде більшою, що дасть змогу краще розпізнавати клієнтів такого класу.

Дякую за увагу!