Тема: Метод логістичної регресії для оцінки кредитоспроможності фізичних осіб

Виконавець роботи:
студент 4 курсу
групи КА-44
Дехтієвський Микола
Дмитрович

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

КИЇВ - 2018

- Мета дослідження
  - полягає у дослідженні та вдосконаленні існуючих методик побудови скорингових моделей та оцінюванні кредитоспроможності фізичних осіб з використанням методу логістичної регресії

- Об’єкт дослідження
  - вибірки статистичних даних, що містять інформацію про надані та непогашені споживчі кредити фізичним особам

- Предмет дослідження
  - методи лінійної ймовірнісної регресії, логістичної регресій та апарат дерев рішень

Актуальність дослідження

- Проблема своєчасного повернення кредитів, що видані фізичним особам, являється актуальною для переважної більшості банківських установ. Вирішення даної проблеми значною мірою являється залежним від «якісної» оцінки кредитоспроможності потенційних клієнтів. Тому саме якісний відбір позичальників, аналіз умов при видачі кредиту, безперервний контроль фінансового стану позичальника та за його здатністю погасити кредит являється однією з основних складових фінансової стабільності кредитних організацій.

Постановка задачі

- виконати аналіз літературних джерел та інформації з Інтернету, щодо існуючих підходів;
- розробити архітектуру системи для оцінювання кредитоспроможності;
- побудувати скорингову модель на основі алгоритму логістичної регресії;
- протестувати модель на реальних даних та провести порівняльний аналіз з іншими методами.

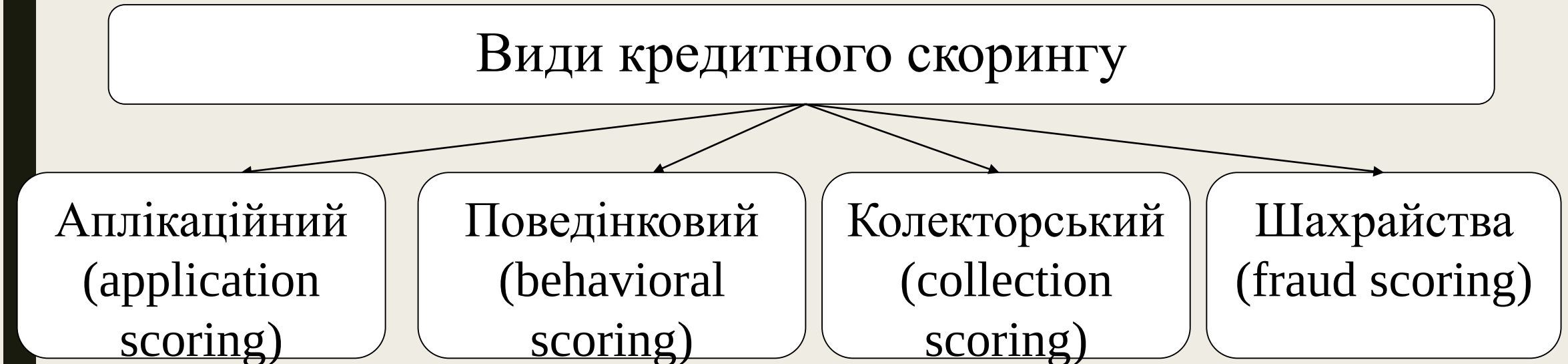
Основні поняття кредитного скорингу

- Оцінка кредитоспроможності фізичної особи – це результати аналізу кількісних показників та якісних характеристик, які можуть бути підтверджені відповідними документами або розрахунками.

- Кредитний скоринг (від англ. score – бал, оцінка), який забезпечує гнучкий інструментарій оцінювання кредитних ризиків і можливість автоматизації процесу прийняття кредитних рішень.

- Основними завданням скорингу є з’ясування спроможності клієнта повернути кредит та з’ясування ступіню надійності та обов'язковості потенційного позичальника. Метою скорингу являється надання реальної бальної оцінки кожному параметру, що характеризує позичальника.

Види кредитного скорингу



Застосування SAS Enterprise Miner в кредитному скорингу

- Всесвітньо визнаним лідером в області інтелектуального аналізу даних та спеціалізованих рішень для вирішення задач скорингу є компанія SAS Institute.

- Enterprise Miner являє собою частину загального сімейства пропонованих SAS систем аналізу даних.

Побудова скорингової моделі
Аналіз і коригування змінних для побудови моделі

- Стратегії підготовки даних для побудови моделі логістичної регресії:
  - *при малій кількості "поганих" позичальників і великому обсязі вибірки взяти 100% всіх «поганих» позичальників і випадковим чином відібрати частину «хороших» спостережень;*
  - *при невеликому обсязі вибірки перезважувати дані, щоб домогтися потрібного співвідношення «поганих» і «хороших» спостережень в вибірці для побудови моделі.*
- Перезважування даних, на відміну від коригування вибірки за рахунок випадкового відбору, дає, як правило, більш надійні оцінки параметрів моделі логістичної регресії

Побудова скорингової моделі

- Логістична регресія - найпоширеніша статистична модель для побудови скорингових карт. Математично модель логістичної регресії виражає залежність логарифма шансу (логіт) від лінійної комбінації незалежних змінних.

$$l\ln(\frac{p_i}{1-p_i}) = b_0 + b_1 x_i^{(1)} + b_2 x_i^{(2)} + \ldots + b_k x_i^{(k)} + \varepsilon_i$$

- де p_i – ймовірність настання дефолту по кредиту для і-го позичальника;

$x_i^{(k)}$ - значення j-ої незалежної змінної;

b_0 - незалежна константа моделі,

b_j – параметри моделі;

ε_i - компонент випадкової помилки.

Логістична регресія

- У разі перезважування вибірки для зміни цього розподілу для більш адекватної подальшої оцінки якості отриманої моделі константу коригують і отримують наступну модель логістичної регресії:

$$ln(\frac{p_i^*}{1-p_i^*}) = ln(\frac{\rho_1\pi_0}{\rho_0\pi_1}) + b_0 + b_1x_i^{(1)} + b_2x_i^{(2)} + \ldots + b_kx_i^{(k)} + \varepsilon_i,$$

- де p_i^* - відкоригована завжди апріорна ймовірність; ρ_1 і ρ_0 - частки «хороших» і «поганих» позичальників у вибірці; π_0 і π_1 - частки «хороших» і «поганих» позичальників в генеральної сукупності.

Логістична регресія

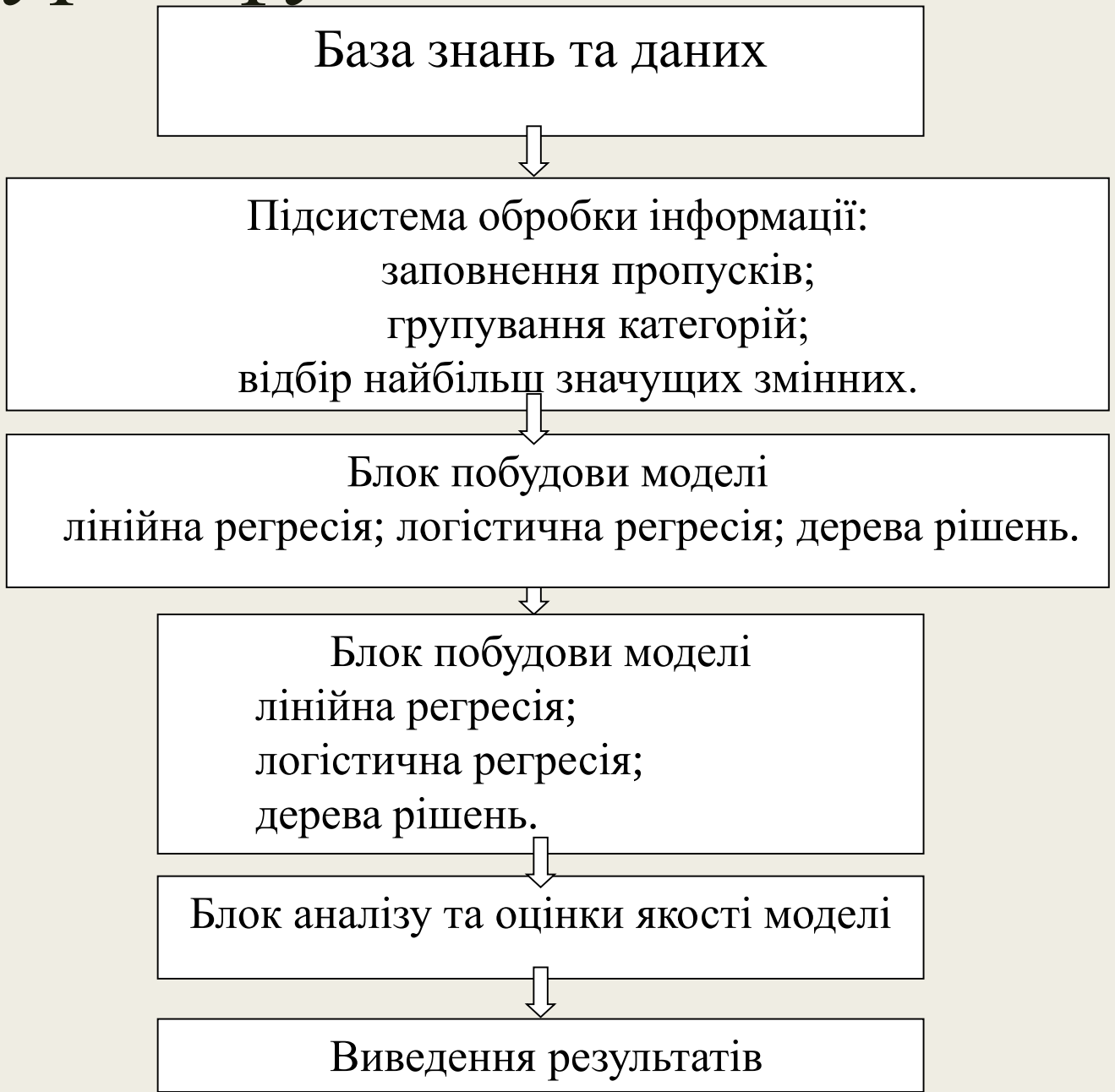
- Для інтерпретації коефіцієнтів моделі логістичної регресії зазвичай використовують експонентну форму запису моделі:

$$p_i = \frac{1}{1 + \exp(-(b_0 + b_1 x_i^{(1)} + b_2 x_i^{(2)} + \ldots + b_k x_i^{(k)} + \varepsilon_i))}.$$

Оцінка якості моделі

- Метод максимальної правдоподібності для оцінки параметрів моделі логістичної регресії.
- Тест Хосмера-Лемешева - додатковий тест для оцінки якості підгонки моделі.
- Статистика Вальда для перевірки гіпотези про значущість окремих коефіцієнтів.
- ROC-крива для оцінки якості класифікації моделі.
- Тест Колмогорова-Смирнова для оцінки якості прогнозування моделі логістичної регресії.

Архітектура і функціональна схема моделі

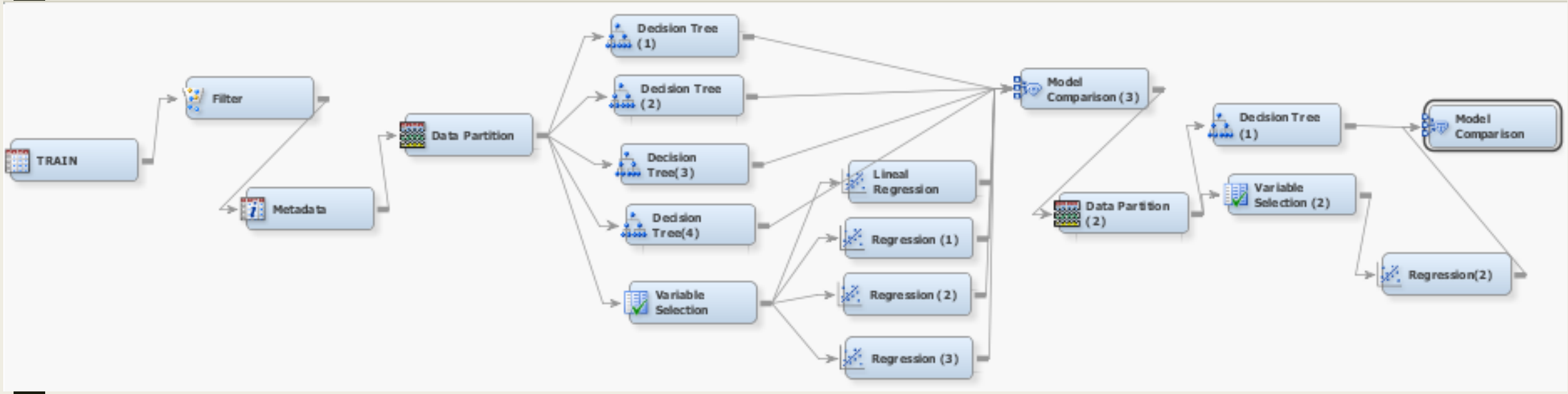


Програмна реалізація

- Для програмної реалізації скориногової системи в рамках дипломної роботи використано програмне середовище системи SAS Enterprise Miner, що має вбудовані засоби для програмування.

- SAS Enterprise Miner обрано як оптимальний інструмент для оцінювання кредитоспроможності фізичних осіб із застосування методу логістичної регресії.

Схема технологічного процесу



Приклад програмного коду SAS Base

```
***  Effect: G_contact ;
_TEMP = 1;
_LP0 = _LP0 + (    0.28713879659879) * _TEMP * _1_0;
 ***  Effect: G_job ;
_TEMP = 1;
_LP0 = _LP0 + (    0.47028211793275) * _TEMP * _2_0;
_LP0 = _LP0 + (   -0.01423042183724) * _TEMP * _2_1;
_LP0 = _LP0 + (   -0.07269205077228) * _TEMP * _2_2;
_LP0 = _LP0 + (   -0.10988831311167) * _TEMP * _2_3;

 ***  Effect: G_month ;
_TEMP = 1;
_LP0 = _LP0 + (      0.810301762588) * _TEMP * _3_0;
_LP0 = _LP0 + (   -0.04116682665246) * _TEMP * _3_1;
_LP0 = _LP0 + (   -0.04116682665246) * _TEMP * _3_1;
_LP0 = _LP0 + (   -0.27766123059281) * _TEMP * _3_2;


***  Effect: balance ;
_TEMP = balance ;
_LP0 = _LP0 + (    0.00003431208538 * _TEMP);
```

```
*** Naive Posterior Probabilities;
drop _MAXP _IY _P0 _P1;
_P0 = probnorm((   -1.53904946696739) +
_LP0);
_P1 = 1.0 - _P0;

REG12DR1:


*** Posterior Probabilities and Predicted Level;
label P_yyes = 'Predicted: y=yes' ;
label P_yno = 'Predicted: y=no' ;
P_yyes = _P0;
_MAXP = _P0;
_IY = 1;
P_yno = _P1;
if (_P1 >  _MAXP + 1E-8) then do;
   _MAXP = _P1;
   _IY = 2;
end;
I_y = REG12DRF[_IY];
U_y = REG12DRU[_IY];
```

Опис набору даних

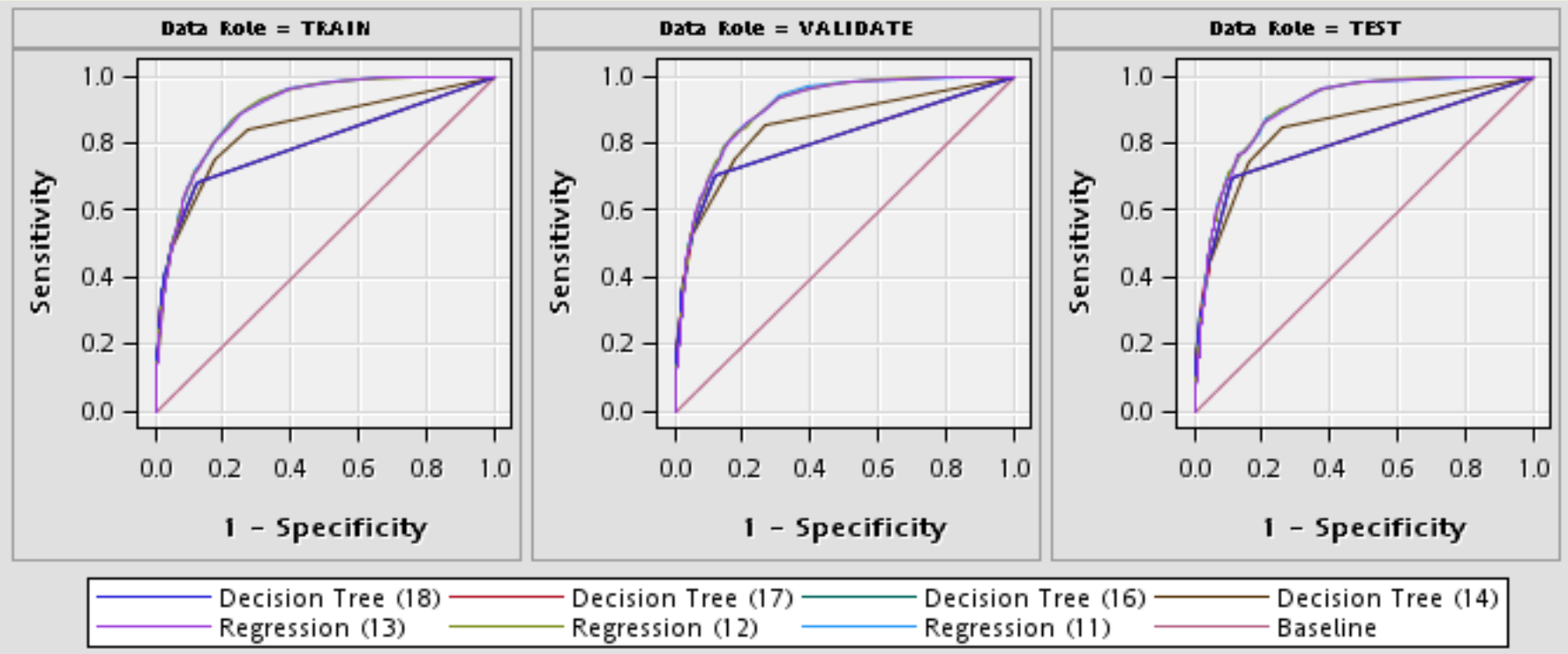
- Був використаний набір даних ООО «Хоум Кредит энд Фінанс Банк», який містив 1048576 записів клієнтів по 35 показникам.
- Отриманий набір даних був розбитий на тренувальну та тестову вибірки у відношенні 80 до 20.

Опис набору даних

- ID - ідентифікатор заявки на кредит;
- SK_DATE_DECISION - дата розгляду заявки;
- DEF - цільова змінна;
- NUM_SOURCE - номер джерела даних;
- CREDIT_ACTIVE - статус кредитного договору;
- CREDIT_COLLATERAL - ознака погашення за рахунок забезпечення;
- CREDIT_CURRENCY валюта кредиту;
- DTIME_CREDIT - дата видачі кредиту;
- CREDIT_DAY_OVERDUE - поточна прострочена заборгованість, дні;
- DTIME_CREDIT_ENDDATE -планова дата закінчення кредиту;
- DTIME_CREDIT_ENDDATE_FACT - фактична дата закінчення кредиту

- Для першої частини аналізу було виконано побудову 8 моделей(4 – дерева рішень, 3 – логістичні регресії, 1 – лінійна регресія).

ROC-криві побудованих моделей



Порівняльна характеристика моделей

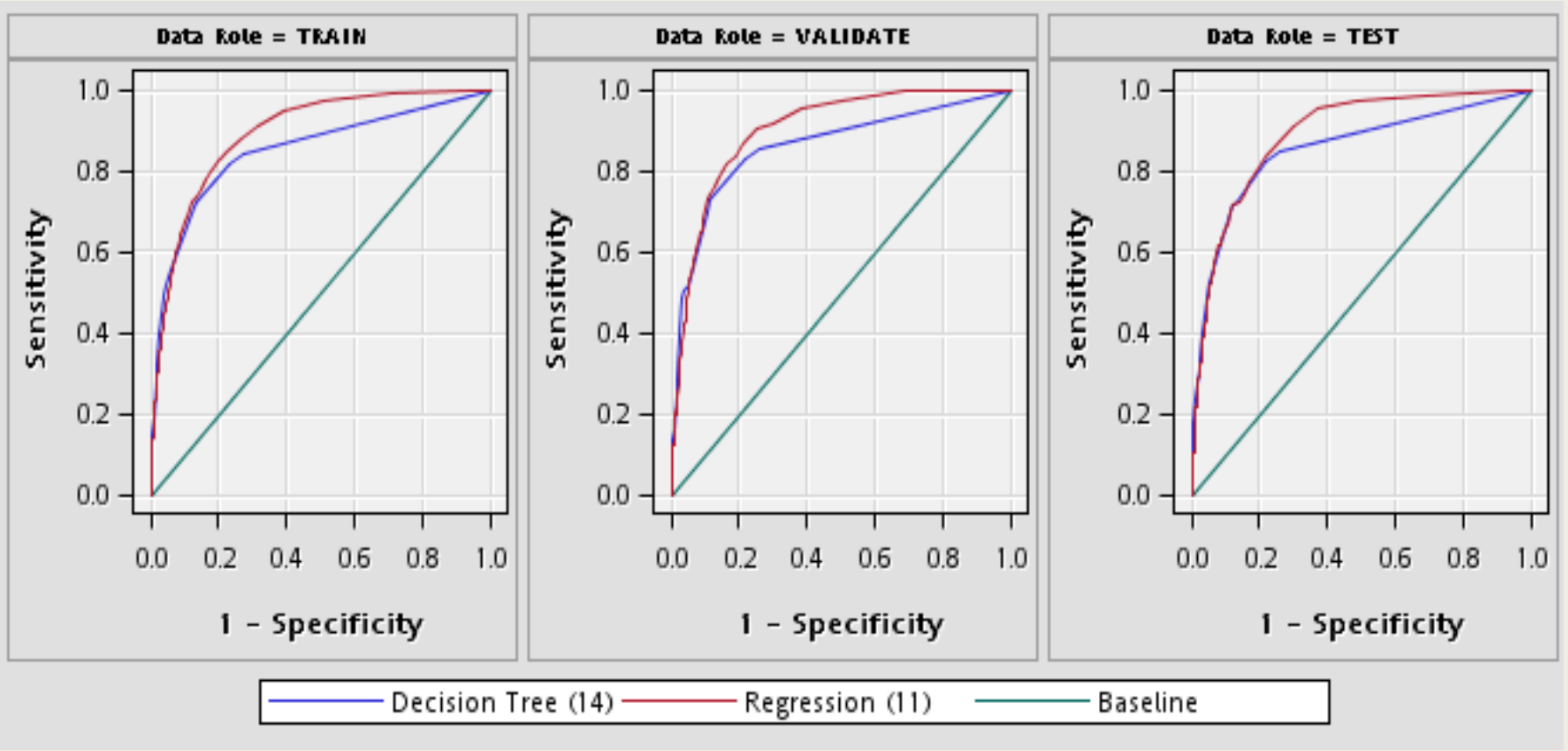
Тип моделі	Навчальна вибірка (Traine)		Тестова вибірка (Test)		Валідаційна вибірка (Validation)	
	ROC індекс	Індекс Gini	ROC індекс	Індекс Gini	ROC індекс	Індекс Gini
Дерево рішень(1)	0,842	0,683	0,844	0,688	0,85	0,7
Дерево рішень(2)	0,799	0,598	0,806	0,611	0,808	0,616
Дерево рішень(3)	0,799	0,598	0,806	0,611	0,808	0,616
Дерево рішень(4)	0,798	0,596	0,805	0,609	0,808	0,616
Лінійна регресія	0,106	0,102	0,105	0,104	0,103	0,105
Логістична регресія(1)	0,899	0,797	0,902	0,803	0,904	0,807
Логістична регресія(2)	0,898	0,797	0,902	0,804	0,904	0,808
Логістична регресія(3)	0,897	0,794	0,901	0,802	0,903	0,807

- З наведених в таблиці результатів для подальшого моделювання були обрані логістична регресія(2) та дерево рішень(1).
- Обрані моделі були створені на інакшому розподілі даних, а саме : 80% навчальна, 10% валідаційна та 10% тестова.

Фінальна порівняльна характеристика моделей

Тип моделі	Навчальна вибірка (Traine)		Тестова вибірка (Test)		Валідаційна вибірка (Validation)	
	ROC індекс	Індекс Gini	ROC індекс	Індекс Gini	ROC індекс	Індекс Gini
Дерево рішень(1)	0,85	0,7	0,851	0,703	0,858	0,717
Логістична регресія(2)	0,892	0,784	0,887	0,775	0,9	0,815

Фінальні графіки ROC-кривих

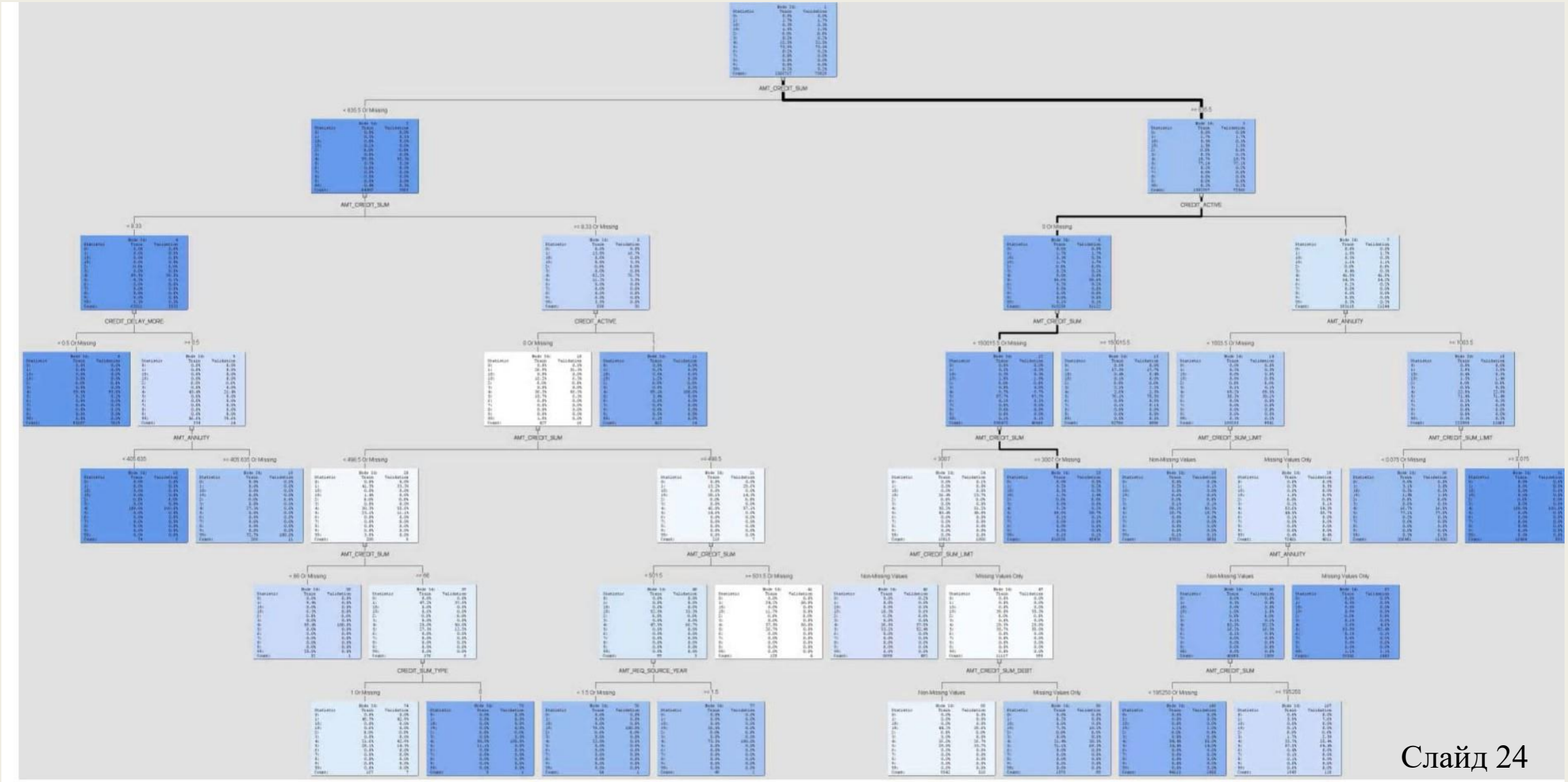


Як видно з порівняльної характеристики моделей та з графіків ROC-кривих за індексом Gini та індексом ROC кращі результати отримали за допомогою методу логістичної регресії.

Статистичні показники побудованої моделі логістичної регресії

Fit Statistics	Statistics Label	Train	Validation	Test
AIC	Информационный критерий Акайка	1968221	.	.
ASE	Средний квадрат ошибки	0.029395	0.029395	0.029436
AVERR	Функция средней ошибки	0.110775	0.110777	0.111047
DFE	Число степеней свободы для ошибки	16400582	.	.
DFM	Число степеней свободы модели	22	.	.
DFT	Общее число степеней свободы	16400604	.	.
DIV	Делитель для асимптотической стандартно	17767321	987077	987415
ERR	Функция ошибки	1968177	109345.7	109649.5
FPE	Итоговая ошибка прогноза	0.029395	.	.
MAX	Максимальная абс. ошибка	1	0.999999	0.999999
MSE	Средний квадрат ошибки	0.029395	0.029395	0.029436
NOBS	Сумма частот	1366717	75929	75955
_NW_	Кол-во оценочных весов	21	.	.
RASE	Квадратный корень из суммы квадратов	0.171449	0.17145	0.17157
RFPE	Корень из итоговой ошибки прогноза	0.171449	.	.
RMSE	Стандартная ошибка	0.171449	0.17145	0.17157
SBC	Байесовский критерий Шварца	1968543	.	.
SSE	Сумма квадратов ошибок	522266.5	29015.06	29065.75
SUMW	Сумма временных частот весов случаев	17767321	987077	987415
MISC	Коэффициент неправильной группировки	0.242244	0.242226	0.242604

Структура дерева рішень.



Висновки

- Запропоновано архітектуру системи для оцінювання кредитоспроможності фізичних осіб;
- Розроблено систему для аналізу, обробки даних, побудови скорингової моделі на основі логістичної регресії для оцінки кредитоспроможності;
- Розроблену систему випробувано на вибірці даних реальних клієнтів банку;
- Виконано порівняльний аналіз з іншими реалізованими методами оцінювання.

Дякую за увагу!