МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Запорізький національний технічний університет

Інформаційні системи та технології в управлінні

МЕТОДИЧНІ ВКАЗІВКИ

теоретичні відомості і завдання до лабораторних робіт

для студентів та магістрів денної форми навчання
спеціальності 7.803060101
Менеджмент організацій і адміністрування
Частина 3
Класифікація у бізнес-аналітиці

2014

Інформаційні системи та технології в управлінні. Методичні вказівки, теоретичні відомості і завдання до лабораторних робіт для студентів та магістрів денної форми навчання спеціальності 7.803060101 Менеджмент організацій і адміністрування. Частина 3. Класифікація в бізнес-аналітиці. / Укл.: Біла Н.І. – Запоріжжя: ЗНТУ, 2014. – с. 50.

Містить теоретичні відомості, індивідуальні завдання та приклади за темою «Класифікація» із курсу «Інформаційні системи та технології в управлінні»

Укладачі: Біла Н.І. доцент,

Рецензенти: Пінчук В.П., доцент
Вишневська В.Г., доцент.

Відповідальний за випуск Корніч Г.В., зав. кафедрою, професор

Затверджено на засіданні кафедри
обчислювальної математики,
протокол № 6 від 28.03.2014

ЗМІСТ

5 ВИРІШЕННЯ ЗАДАЧ КЛАСИФІКАЦІЇ

5.1 Опис процесу класифікації

Задача класифікації – це задача розбиття множини об'єктів або спостережень на апріорно задані групи, називані класами, всередині кожної з яких вони вважаються схожими один на одного, та мають приблизно однакові властивості й ознаки. При цьому рішення здійснюється на основі аналізу значень атрибутів (ознак).

Класифікація є однією з найважливіших задач Data Mining. Вона застосовується в маркетингу при оцінці кредитоспроможності позичальників, визначенні лояльності клієнтів, розпізнаванні образів, медичній діагностиці й багатьох інших сферах. Якщо аналітикові відомі властивості об'єктів кожного класу, то коли нове спостереження відноситься до певного класу, дані властивості автоматично поширюються й на нього.

Якщо число класів обмежено двома, то має місце бінарна класифікація, до якої можуть бути зведені багато більш складних задач. Наприклад, замість визначення таких ступенів кредитного ризику, як «Високий», «Середній» або «Низький», можна використовувати всього дві - «Видати» або «Відмовити».

Для проведення класифікації за допомогою математичних методів необхідно мати формальний опис об'єкта, яким можна оперувати, використовуючи математичний апарат класифікації. Таким описом найчастіше виступає база даних. Кожний запис бази даних несе інформацію про деяку властивість об'єкта.

Набір вхідних даних (вибірку даних) розбивають на дві множини: навчальна (training set) і тестова (test set).

У навчальну вибірку входять об'єкти, для яких відомі значення як незалежних, так і залежних змінних. На підставі навчальної вибірки будується модель визначення значення залежної змінної. Її часто називають функцією класифікації. Для одержання максимально точної функції до навчальної вибірки пред'являються такі основні вимоги:

- кількість об'єктів, які входять у вибірку, повинне бути досить великим. Чим більше об'єктів, тім побудована на їхній основі функція класифікації буде точніше;
- у вибірку повинні входити об'єкти, які представляють усі можливі класи;
- для кожного класу вибірка повинна мати достатню кількість об'єктів.

Тестова (test set) множина також містить вхідні й вихідні значення параметрів. Тут вихідні значення використовуються для перевірки працездатності моделі.

Процес класифікації складається із двох етапів: конструювання моделі і її використання.

a) Конструювання моделі здійснюється на основі навчальної вибірки. В результаті одержуємо модель, яка представляється або класифікаційними правилами або деревом розв'язків або математичною формулою або комп'ютерним об'єктом (як нейронні мережі).

b) Використання моделі: класифікація нових або невідомих значень.

   1) Оцінка правильності (точності) моделі. Відомі значення з тестового набору порівнюються з результатами використання отриманої моделі. За рівень точності ухвалюється відсоток правильно класифікованих прикладів у тестовій множині.

   2) Якщо точність моделі припустима, можливе використання моделі для класифікації нових прикладів, клас яких невідомий.

Основні проблеми, з якими зустрічаються при розв'язку задач класифікації, - це незадовільна якість вхідних даних, у яких зустрічаються як помилкові дані, так і пропущені значення, різні типи атрибутів - числові й категоріальні, різна значимість атрибутів, а також так звані проблеми overfitting і underfltting. Суть першої з них полягає в тому, що класифікаційна функція при побудові "занадто добре" адаптується до даних, і помилки, які зустрічаються в них, і аномальні значення намагається інтерпретувати як частину внутрішньої структури даних. Очевидно, що така модель буде

некоректно працювати надалі з іншими даними, де характер помилок буде дещо іншим. Ця ситуація проявляється в тому, що помилка на тестовій множині значно більше помилки на навчальній. Терміном underfltting позначають ситуацію, коли спостерігається занадто велика кількість помилок при перевірці класифікатора на навчальній множині. Це означає, що особливих закономірностей у даних не було виявлено й або їх немає взагалі, або необхідно вибрати інший метод їх виявлення.

5.2 Методи, застосовувані для вирішення задач класифікації

Методи умовно поділяють на дві групи.

1) Статистичні методи класифікації

- байєсовська (наївна) класифікація;
- логістична регресія;
- дискримінантний аналіз

2) Методи машинного навчання

- класифікація за допомогою дерев рішень;
- класифікація за допомогою штучних нейронних мереж;
- класифікація за допомогою алгоритмів покриття;
- класифікація методом опорних векторів;
- класифікація за допомогою методу k-найближчих сусідів;
- класифікація CBR-методом.

5.3 Геометрична інтерпретація задачі класифікації

Задача класифікації має геометричну інтерпретацію. Розглянемо її на прикладі із двома незалежними змінними, що дозволить представити її у двовимірному просторі (рис. 1.1). Кожному об'єкту ставитися у відповідність точка на площині. Символи "+" і "-" позначають приналежність об'єкта до одного з двох класів. Очевидно, що дані мають чітко виражену структуру: усі точки класу "+" зосереджені в центральній області. Побудова класифікаційної функції зводиться до побудови поверхні, що обводить центральну область. Вона визначається як функція, що має значення "+" усередині обведеної області й "-" — поза нею.

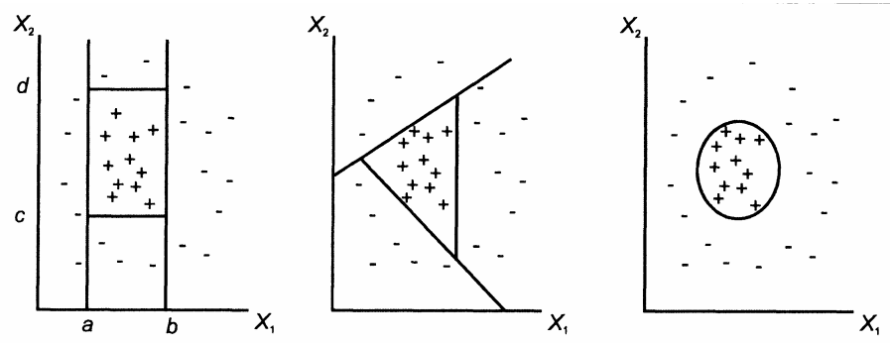


Рисунок 5.1 - Класифікація у двовимірному просторі

Як видно з рис. 5.1, є кілька можливостей для побудови такої області. Вид функції залежить від застосовуваного алгоритму.

5.4 Оцінка якості моделі класифікації

Оцінка точності класифікації може проводитися за допомогою крос-перевірки (cross-validation), що описана в п.1.5, або за допомогою тестової множини. Точність класифікації тестової множини порівнюється з точністю класифікації навчальної множини. Якщо класифікація тестової множини дає приблизно такі ж результати по точності, як і класифікація навчальної множини, вважається, що дана модель пройшла перевірку.

Для візуалізації результатів перевірки найчастіше використовують таблицю спряженості. Розглянемо докладніше процес її створення. Для вирішення задачі класифікації використовується таблиця, у якій уже є вихідний стовпець, що містить клас об'єкта. Після застосування навчального алгоритму додається ще один стовпець із вихідним полем, але його значення вже обчислюються, використовуючи побудовану модель. При цьому значення в стовпцях можуть відрізнятися. Чим більше таких відмінностей, тим гірше побудована модель класифікації.

Зовнішній вигляд таблиці спряженості наведений на рисунку 5.2.

Нижче представлене пояснення щодо складових таблиці.

TP (True Positives) - кількість вірно класифікованих позитивних прикладів (так звані істинно позитивні випадки).

	Классифицировано		
Фактически	Да	Нет	Итого
Да	TP	FN	
Нет	FP	TN	
Итого			

Рисунок 5.2 - Таблиця спряженості

TN (True Negatives) - кількість вірно класифікованих негативних прикладів (істинно негативні випадки).

FN (False Negatives) - кількість позитивних прикладів, класифікованих як негативні (помилка I роду). Це так званий «помилковий пропуск», коли подія, що нас цікавить, помилково не виявляється (хибно негативні приклади).

FP (False Positives) - негативні приклади, класифіковані як позитивні (помилка II роду). Це «помилкове виявлення», тому що при відсутності події помилково виноситься рішення про її присутність (хибно позитивні випадки).

Що є позитивною подією, а що - негативною, залежить від конкретної задачі.

На головній діагоналі показана кількість правильно класифікованих прикладів, на побічній діагоналі – кількість неправильно класифікованих прикладів.

Якщо кількість неправильно класифікованих прикладів досить велика, це говорить про погано побудовану модель, потрібно змінити параметри побудови моделі, збільшити навчальну вибірку або змінити набір вхідних полів. Якщо ж кількість неправильно класифікованих прикладів незначна, це може говорити про те, що дані приклади є аномаліями. У цьому випадку можна подивитися, чим же характеризуються такі приклади й, можливо, додати новий клас для їхньої класифікації.

Ще один засіб, який застосовується для подання та оцінки результатів бінарної класифікації в машинному навчанні - ROC-аналіз.

ROC-аналіз дозволяє провести оцінку якості моделі класифікатора, порівняти прогностичну силу декількох моделей, визначити оптимальну точку відсікання для віднесення об'єктів до того чи іншого класу. При цьому передбачається, що у класифікатора є додаткові параметри, що дозволяють вже після проведеного навчання варіювати співвідношення помилок першого й другого роду.

В основі ROC-аналізу лежить побудова графіків - ROC-кривих (Receiver Operator Characteristic). Назва прийшла із систем обробки сигналів. Оскільки класів два, один з них називається класом з позитивними наслідками, другий - з негативними. ROC-крива показує залежність кількості вірно класифікованих позитивних прикладів від кількості невірно класифікованих негативних прикладів. У термінології ROC-аналізу перші називаються істинно позитивною, другі - хибно негативною множиною. Як уже говорилося вище, у класифікатора є деякий параметр, варіюючи який, ми будемо одержувати ту або іншу розбивку на два класи. Цей параметр часто називають порогом, або точкою відсікання (cutoff value).

При аналізі використовуються значення з таблиці спряженості, але найчастіше оперують не абсолютними показниками, а відносними - частками (rates) вираженими у відсотках:

Так, частка істинно позитивних прикладів (True Positives Rate):

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \cdot 100\% \cdot \tag{5.1}$$

Частка хибно позитивних прикладів (False Positives Rate):

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}} \cdot 100\% \cdot \tag{5.2}$$

Введемо ще два визначення: чутливість і специфічність моделі. Ними визначається об'єктивна цінність будь-якого бінарного класифікатора.

Чутливість (Sensitivity) - це і є частка істинно позитивних випадків:

$$\mathrm{Se} = \mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \cdot 100\% \cdot \tag{5.3}$$

Специфічність (Specificity) - частка істинно негативних випадків, які були правильно ідентифіковані моделлю:

$$\mathrm{Sp} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}} \cdot 100\% \cdot \tag{5.4}$$

Зауважимо, що $\mathrm{FPR} = 100 - \mathrm{Sp}.$ (5.5)

Модель із високою чутливістю часто дає істинний результат при наявності позитивного результату (виявляє позитивні приклади). Навпаки, модель із високою специфічністю частіше дає істинний результат при наявності негативного результату (виявляє негативні приклади).

ROC-крива будується у такий спосіб: для кожного значення порога відсікання, що змінюється від 0 до 1 із кроком dx (наприклад, 0.01) розраховуються значення чутливості Se і специфічності Sp. Як альтернатива порогом може бути кожне наступне значення приклада у вибірці. Будується графік залежності: по осі Y відкладається чутливість Se, по осі X - 100%-Sp (сто відсотків мінус специфічність).

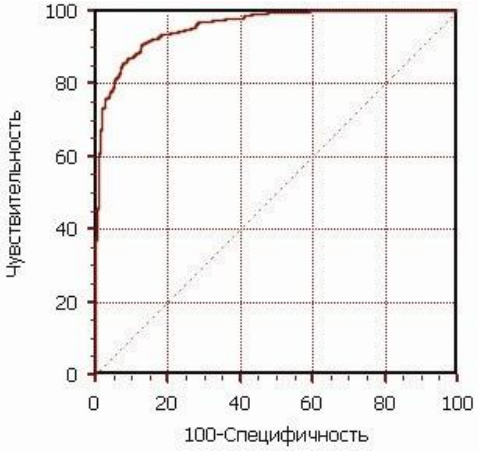


Рисунок 5.3 - ROC-крива

Графік часто доповнюють прямою y=x.

Для ідеального класифікатора графік ROC-кривої проходить через верхній лівий кут, де частка істинно позитивних випадків становить 100% або 1.0 (ідеальна чутливість), а частка хибно позитивних

прикладів дорівнює нулю. Тому чим ближче крива до верхнього лівого кута, тим вища прогностична здатність моделі. Навпаки, чим менше вигин кривої й чим ближче вона розташована до діагональної прямої, тим менш ефективна модель. Діагональна лінія відповідає "марному" класифікатору, тобто повної нерозрізненості двох класів.

При візуальній оцінці ROC-кривих розташування їх відносно один одного вказує на їхню порівняльну ефективність. Крива, розташована вище й лівіше, свідчить про більшу прогностичну здатність моделі. Так, на рисунку 5.4 дві ROC-криві сполучені на одному графіку. Видно, що модель "А" краща.

Візуальне порівняння кривих ROC не завжди дозволяє виявити найбільш ефективну модель. Своєрідним методом порівняння ROC-кривих є оцінка площі під кривими. Теоретично вона змінюється від 0 до 1.0, але, через те, що модель завжди характеризується кривою, розташованою вище позитивної діагоналі, то звичайно говорять про зміни від 0.5 ("марний" класифікатор) до 1.0 ("ідеальна" модель). Ця оцінка може бути отримана безпосередньо обчисленням площі під багатогранником, обмеженим праворуч і знизу осями координат і ліворуч угорі - експериментально отриманими точками (рис. 5.3).

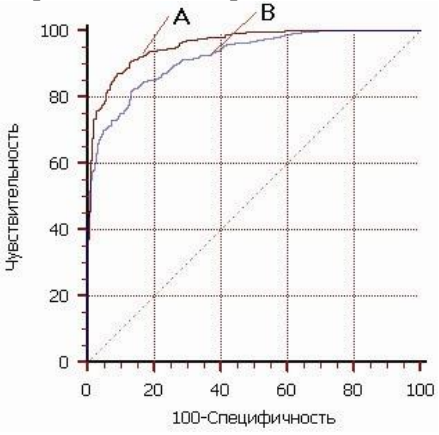


Рисунок 5.4 - Порівняння ROC-кривих

Чисельний показник площі під кривою називається AUC (Area Under Curve). Обчислити його можна, наприклад, за допомогою чисельного методу трапецій:

$$AUC = \int f(x)dx = \sum_i \left[\frac{x_{i+1} + x_i}{2} \right] \times (Y_{i+1} - Y_i). \quad (5.6)$$

З великими допущеннями можна вважати, що чим більше показник AUC, тим кращою прогностичною силою володіє модель. Однак варто знати, що:

- показник AUC призначений скоріше для порівняльного аналізу декількох моделей;
- AUC не містить ніякої інформації про чутливість і специфічність моделі.

У літературі іноді приводиться експертна шкала для значень AUC - таблиця 1.1, по якій можна судити про якість моделі:

Ідеальна модель володіє 100% чутливістю й специфічністю. Однак на практиці домогтися цього неможливо, більше того, неможливо одночасно підвищити й чутливість, і специфічність моделі. Компроміс знаходиться за допомогою порога відсікання, тому що його граничне значення впливає на співвідношення Se і Sp. Можна говорити про задачу знаходження оптимального порога відсікання (optimal cutoff value).

Таблиця 5.1 - Оцінка якості класифікації на підставі AUC

Інтервал AUC	**Якість моделі**
0.9 – 1.0	Відмінна
0.8 – 0.9	Дуже добра
0.7 – 0.8	Добра
0.6 – 0.7	Середня
0.5 – 0.6	Незадовільна

Поріг відсікання потрібний для того, щоб застосовувати модель на практиці: відносити нові приклади до одного із двох класів. Для визначення оптимального порога потрібно задати критерій його визначення, тому що в різних задачах присутня своя оптимальна стратегія. Критеріями вибору порога відсікання можуть виступати:

- Вимога максимальної сумарної чутливості й специфічності моделі, тобто: $Cutoff = \max_k (Se_k + Sp_k)$

- Вимога балансу між чутливістю й специфічністю, тобто коли $\mathrm{Se} \approx \mathrm{Sp}: \ \mathrm{Cutoff} = \min_{k} \left| \mathrm{Se}_k - \mathrm{Sp}_k \right|$

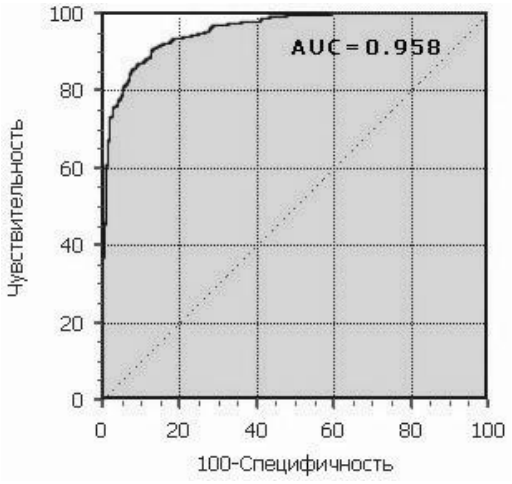


Рисунок 5.5 - Площа під ROC-кривою

У другому випадку порогом є точка перетинання двох кривих, коли по осі X відкладається поріг відсікання, а по осі Y - чутливість і специфічність моделі (рис. 1.6).

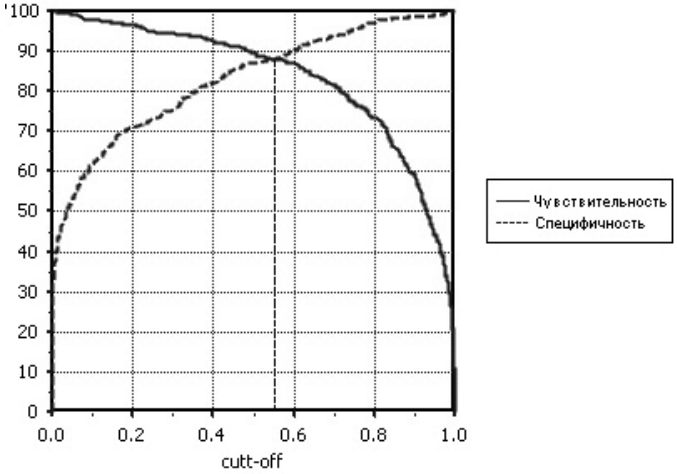


Рисунок 5.6 - "Точка балансу" між чутливістю й специфічністю

5.5 Методи класифікації

1) Логістична регресия

Логістична регресія застосовується для пророкування ймовірності виникнення деякої події за значеннями множини ознак. Для цього вводиться так звана *залежна змінна y*, що приймає лише одне із двох значень — як правило, це числа 0 (подія не відбулася) і 1 (подія відбулася), і множина *незалежних змінних* (також називаних ознаками, предикторами або регресорами) — дійсних чисел $x_1, x_2, ..., x_n$, на основі значень яких потрібно обчислити ймовірність прийняття того або іншого значення залежної змінної.

На практиці логістична регресія використовується для розв'язку задач класифікації з лінійно-поділюваними класами.

Задана вибірка – множина m пар (x_i, y_i), у яких опис *i*-го елемента $x_i \in \Re^n$, і значення залежної змінної $y \in \{0,1\}$.

Прийнята модель логістичної регресії, згідно з якою вільні змінні X й залежна змінна *y* зв'язані залежністю

$$y = f(b,x) = \frac{1}{1+e^{-z}} + \varepsilon \qquad \text{де} \quad z = b_0 + \sum_{j=1}^{n} b_j x_j \ .$$

Приймемо позначення $p_i = f(\mathbf{b}, x_i)$, вектор $\mathbf{b} = [b_0, b_1, .. b_n]^T$.

Потрібно знайти таке значення вектора параметрів $\mathbf{b}$, яке б доставляло мінімум нормі вектора нев'язок

$$S = |\mathbf{y} - \mathbf{p}|^2 = \sum_{i=1}^{m} (y_i - p_i)^2$$

Оптимальні параметри відшукуються послідовно за допомогою ітераційного методу найменших квадратів.

Насправді, логістичну регресію можна представити у вигляді одношарової нейронної мережі із сигмоїдальною функцією активації, ваги якої є коефіцієнти логістичної регресії, а вага поляризації – константа регресійного рівняння (рис. 5.7).

Як відомо, одношарова нейронна мережа може успішно розв'язати лише задачу лінійної сепарації. Тому можливості по моделюванню нелінійних залежностей у логістичної регресії відсутні. Однак для оцінки якості моделі логістичної регресії існує ефективний інструмент ROC-аналіз, що є безсумнівною його перевагою.

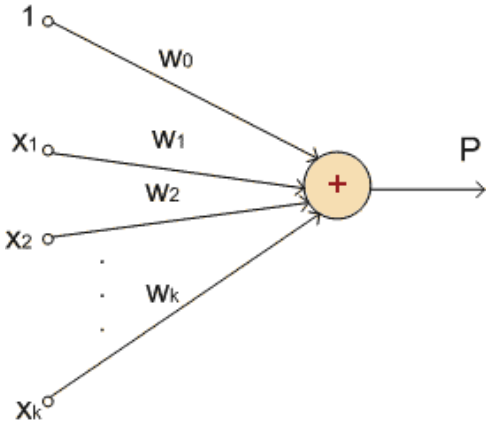


Рисунок 5.7 – Представлення логістичної регресії у вигляді нейронної мережі

2) Дерева рішень

Дерева рішень належать до самих популярних і потужних інструментів Data Mining, що дозволяють ефективно вирішувати задачі класифікації. В основі роботи дерев рішень лежить процес рекурсивної розбивки вхідної множини спостережень або об'єктів на підмножини, асоційовані із класами.

Процес конструювання дерева рішень. Нагадаємо, що розглянута нами задача класифікації відноситься до стратегії навчання із учителем. У цих випадках усі об'єкти тренувального набору даних заздалегідь віднесені до одного з визначених класів.

Алгоритми конструювання дерев рішень складаються з етапів "побудова" або "створення" дерева (tree building) і "скорочення" дерева (tree pruning). У ході створення дерева вирішуються питання вибору критерію розщеплення й зупинки навчання (якщо це

передбачене алгоритмом). У ході етапу скорочення дерева вирішується питання відсікання деяких його гілок.

Критерій розщеплення. Процес створення дерева відбувається зверху вниз, тобто є спадним. У ході процесу алгоритм повинен знайти такий критерій розщеплення, іноді також називаний критерієм розбивки, щоб розбити множину на підмножини, які б асоціювалися з даним вузлом перевірки. Кожний вузол перевірки повинен бути позначений певним атрибутом. Існує правило вибору атрибута: він повинен розбивати вхідну множину даних таким чином, щоб об'єкти підмножин, одержуваних у результаті цієї розбивки, були представниками одного класу або ж були максимально наближені до такої розбивки. Остання фраза означає, що кількість об'єктів з інших класів, так званих "домішок", у кожному класі повинне прагнути до мінімуму.

Існують різні критерії розщеплення. Найбільш відомі - міра ентропії й індекс Gini. Якщо задана множина *T*, що включає приклади з *n* класів, індекс Gini, визначається по формулі:

$$gini(T) = 1 - \sum_{j=1}^{n} p_j^2, \qquad p_j = \frac{N_j}{N} \text{ - частка класу } j \text{ у вузлі } T.$$

де T - поточний вузол, p_j - імовірність класу j у вузлі *T*, *n* - кількість класів, N – кількість об'єктів у вузлі.

Міра ентропії при побудові дерев рішень – це міра різноманітності класів у вузлі. У результаті розбивки повинні утворюватися вузли з меншою різноманітністю станів вихідної змінної. Отже, ентропія падає, а кількість внутрішньої інформації у вузлі росте. Формально ентропія певного вузла T дерева рішень визначається:

$$Info(T) = -\sum_{j=1}^{n} p_j \cdot \log(p_j).$$

Ентропія всієї розбивки – сума ентропій усіх вузлів, помножена на частку записів кожного вузла в загальнім числі записів:

$$Info(S) = \frac{N_1}{N} Info(T_1) + \frac{N_2}{N} Info(T_2) + \ldots + \frac{N_n}{N} Info(T_n).$$

Для вибору атрибута розщеплення використовується критерій, що називається приростом інформації або зменшенням ентропії:

$$Gain(S) = Info(T) - Info_s(T).$$

У якості найкращого атрибута для використання в розбивці S вибирається той, який забезпечує найбільший приріст інформації *Gain(S).*

Велике дерево не означає, що воно "підходяще". Чим більше окремих випадків описане в дереві рішень, тем менша кількість об'єктів попадає в кожний окремий випадок. Такі дерева називають "гіллястими" або "рунистими", вони складаються з невиправдано великої кількості вузлів і гілок, вихідна множина розбивається на велику кількість підмножин, що випливає із дуже малого числа об'єктів. У результаті "переповнення" таких дерев їх здатність до узагальнення зменшується, і побудовані моделі не можуть давати вірні відповіді.

У процесі побудови дерева, щоб його розміри не стали надмірно великими, використовують спеціальні процедури, які дозволяють створювати оптимальні дерева, так звані дерева "підходящих розмірів" (Breiman,1984).

Який розмір дерева може вважатися оптимальним? Дерево повинне бути досить складним, щоб враховувати інформацію з досліджуваного набору даних, але одночасно воно повинне бути досить простим. Інакше кажучи, дерево повинне використовувати інформацію, що поліпшує якість моделі, і ігнорувати ту інформацію, яка її не поліпшує.

Отут існує дві можливі стратегії. Перша полягає в нарощуванні дерева до певного розміру відповідно до параметрів, заданих користувачем.

Визначення цих параметрів може ґрунтуватися на досвіді й інтуїції аналітика, а також на деяких "діагностичних повідомленнях" системи, що конструює дерево рішень.

Друга стратегія полягає у використанні набору процедур, що визначають "підходящий розмір" дерева, вони розроблені Бриманом,

Куілендом і ін. в 1984 році. Однак, як відзначають автори, не можна сказати, що ці процедури доступні починаючому користувачеві.

Процедури, які використовують для запобігання створення надмірно великих дерев, включають: скорочення дерева шляхом відсікання гілок; використання правил зупинки навчання.

Слід зазначити, що не всі алгоритми при конструюванні дерева працюють по одній схемі. Деякі алгоритми включають два окремі послідовні етапи: побудова дерева і його скорочення; інші чергують ці етапи в процесі своєї роботи для запобігання нарощування внутрішніх вузлів.

Зупинка побудови дерева. Розглянемо правило зупинки. Воно повинне визначити, чи є розглянутий вузол внутрішнім вузлом, при цьому він буде розбиватися далі, або ж він є кінцевим вузлом, тобто вузлом розв'язком.

Зупинка - такий момент у процесі побудови дерева, коли слід припинити подальші розгалуження.

Один з варіантів правил зупинки - "рання зупинка" (prepruning), вона визначає доцільність розбивки вузла. Перевага використання такого варіанта - зменшення часу на навчання моделі. Однак тут виникає ризик зниження точності класифікації. Тому рекомендується " замість зупинки використовувати відсікання" (Breiman, 1984).

Другий варіант зупинки навчання - обмеження глибини дерева. У цьому випадку побудова закінчується, якщо досягнута задана глибина.

Ще один варіант зупинки - завдання мінімальної кількості прикладів, які будуть утримуватися в кінцевих вузлах дерева. При цьому варіанті розгалуження тривають до того моменту, поки всі кінцеві вузли дерева не будуть чистими або будуть містити не більш ніж задане число об'єктів.

Скорочення дерева або відсікання гілок. Розв'язком проблеми занадто гіллястого дерева є його скорочення шляхом відсікання (pruning) деяких гілок.

Якість класифікаційної моделі, побудованої за допомогою дерева рішень, характеризується двома основними ознаками: точністю розпізнавання й помилкою.

Відсікання гілок або заміну деяких гілок піддеревом слід проводити там, де ця процедура не приводить до зростання помилки. Процес проходить знизу нагору, тобто є висхідним. Це більш популярна процедура, чим використання правил зупинки. Дерева, одержувані після відсікання деяких гілок, називають усіченими.

Якщо таке усічене дерево усе ще не є інтуїтивним і складно для розуміння, використовують добування правил, які поєднують у набори для опису класів. Кожний шлях від кореня дерева до його вершини або листа дає одне правило. Умовами правила є перевірки на внутрішніх вузлах дерева.

Алгоритми. На сьогоднішній день існує велика кількість алгоритмів, що реалізують дерева рішень: CART, C4.5, CHAID, CN2, Newid, Itrule і інші.

Алгоритм C 4.5. Алгоритм C4.5 будує дерево рішень з необмеженою кількістю гілок у вузлах. Даний алгоритм може працювати тільки з дискретним залежним атрибутом і тому може вирішувати тільки задачі класифікації. C4.5 вважається одним з найвідоміших і широко використовуваних алгоритмів побудови дерев класифікації. Алгоритм C4.5 повільно працює на надвеликих наборах даних й таких, що мають багато шумів.

Основні характеристики алгоритму CART: бінарне розщеплення, критерій розщеплення - індекс Gini, алгоритми minimal cost-complexity tree pruning і V-fold cross-validation, принцип "виростити дерево, а потім скоротити", висока швидкість побудови, обробка пропущених значень.

Алгоритми побудови дерев рішень різняться наступними характеристиками:

- вид розщеплення - бінарне (binary), множинне (multi-way);
- критерії розщеплення - ентропія, Gini, інші;
- можливість обробки пропущених значень;
- процедура скорочення гілок або відсікання;
- можливості добування правил з дерев.

Жоден алгоритм побудови дерева не можна апріорі вважати найкращим або довершеним, підтвердження доцільності використання конкретного алгоритму повинне бути перевірене й підтверджене експериментом.

5.6 Приклад вирішення задачі класифікації на основі логістичної регресії

Скорингові моделі для оцінки кредитоспроможності позичальників. Технологіям кредитного скоринга – автоматичної оцінки кредитоспроможності фізичної особи – у банківській сфері традиційно приділяється підвищена увага. Сьогодні можна сказати, що експертні методи йдуть у минуле, і все частіше при розробці скорингових моделей звертаються до алгоритмів Data Mining. Класичну скорингову карту можна побудувати за допомогою логістичної регресії на основі накопиченої кредитної історії, застосувавши до неї Roc-Аналіз для керування ризиками. Крім того, гарні моделі, що легко інтерпретуються, можна одержати, використовуючи дерева рішень.

5.6.1 Постановка задачі

У комерційному банку є продукт «Нецільовий споживчий кредит»: кредити надаються на будь-які потреби з ухваленням рішення протягом декількох годин. За цей час перевіряються мінімальні відомості про клієнта, в основному такі, як відсутність кримінального минулого й кредитна історія в інших банках.

У банку накопичена статистична інформація про позичальників і якість обслуговування ними боргу за кілька місяців. Керівництво банку, розуміючи, що відсутність адекватних математичних інструментів, що дозволяють оптимізувати ризики, не сприяє розширенню роздрібного бізнесу в області споживчого кредитування, поставило перед відділом роздрібних ризиків задачу розробити скорингові моделі з різними стратегіями кредитування, які дозволили б управляти ризиками, підбираючи рівень схвалень кредитів, і мінімізувати число безнадійних позичальників.

Вхідні дані. Інформація про позичальників – фізичних особах і кредитних договорах зберігається в банківській інформаційній системі. Там же зберігаються графіки й дати погашень кредиту,

відомості про прострочення, про їхні суми, про відсотки і т.д. Будемо вважати, що ми одержали цю інформацію у вигляді текстового файлу.

Важливим є питання, що розуміти під параметрами позичальника. У банківській практиці перед скорингом позичальник, як правило, проходить процедуру андеррайтингу – перевірку на задоволення жорстким вимогам: відповідність віку відсутність кримінального минулого, наявність певного доходу. При цьому висуваються вимоги до мінімального рівня доходу й розраховується можливий ліміт кредиту. При його розрахунках бере участь один із двох коефіцієнтів - або П/Д.

Коефіцієнт «Платіж/Дохід» (П/Д) – відношення щомісячних платежів по кредиту позичальника до його доходу за цей період. Уважається, що значна величина цього коефіцієнта свідчить про підвищений ризик як для кредитора так і для позичальника.

Коефіцієнт «Зобов'язання/Дохід» (З/Д) – відношення щомісячних зобов'язань позичальника до його доходу за той же період з урахуванням утримань податків. У зобов'язання включаються витрати, пов'язані з виплатою планованого кредиту, а також наявні інші довгострокові зобов'язання. Вважається, що розмір щомісячних зобов'язань позичальника не повинен перевищувати 50-60% його сукупного чистого доходу.

Заявки клієнтів, що не пройшли андеррайтинг, одержать відмову й навіть не потраплять на скоринг. Тому на вхід скорингової процедури вигідніше подавати не дохід, а відношення П/Д або З/Д.

У нашій задачі представлено 2709 кредитів (файл loans.txt) з відомими результатами платежів протягом декількох місяців після видачі кредиту. Набір даних уже розбитий на дві множини – навчальна (80%) і тестова (20%) так, щоб у кожній множині частка поганих кредитів була приблизно однакова. Структура й опис полів текстового файлу із кредитними історіями наведено в таблиці 5.2.

Таблиця 5.2 - Дані по позичальниках

№	Поле	Опис	Тип
1	Код	Службовий код заявки	Цілий
2	Дата	Дата видачі кредиту	Дата/час
3	З/Д %	Коефіцієнт Зобов'язання/Дохід	Дійсний

4	Вік	Вік позичальника на момент ухвалення рішення про видачу кредиту	Цілий
5	Проживання	Підстава для проживання: власник; муніципальне житло; оренда.	Строковий
6	Строк проживання у регіоні	Менше 1 року; від 1 року до 5 років; понад 5 років.	Строковий
7	Сімейне положення	неодружений/ незаміжня; одружений/замужем; розведений(-а)/вдівство; інше.	Строковий
8	Освіта	Середнє; середнє спеціальне; вище.	Строковий
9	Стаж роботи на останньому місці	Менш 1 року; від 1 року до 3 років; понад 3 років.	Строковий
10	Рівень посади	Співробітник; керівник середньої ланки; керівник вищої ланки	Строковий
11	Кредитна історія	Інформація береться з бюро кредитних історій. Якщо є негативна інформація про клієнта (прострочення по минулих кредитах), то йому привласнюється категорія «негативна»	Строковий
12	Прострочення понад 60 дні	Факт наявності прострочень понад 60 дні: 0 – були відсутні; 1 – мали місце	Цілий
13	Тестова множина	Службова ознака, відповідальна за те, до якої множини відноситься запис. TRUE відповідає тестовій множині.	Логічний

5.6.2 Скорингова карта на основі логістичної регресії

Базовим статистичним алгоритмом, який будує аналог скорингової карти, є логістична регресія.

Розглянемо поняття «шанс», яке визначається як імовірність того, що подія відбулася (шанс успіху), розділена на ймовірність того, що подія не відбулася (шанс неуспіху). Шанси й імовірності містять ту саму інформацію, але по-різному її виражають. Якщо ймовірність того, що подія відбудеться, позначити ρ, то шанси цієї події дорівнюють $\rho/(1-\rho)$. Наприклад, якщо ймовірність видужання становить 0,3, то шанси видужати дорівнюють 0,3/(1-0,3)=0,43.

Визначимо також так зване *відношення шансів*, або *відношення незгоди* (odds ratio – OR), що є відношенням шансів того, що подія відбудеться, до шансів того, що подія не відбудеться. Очевидно, що якщо OR=1, то модель виявляється марною, і чим сильніше відношення шансів відрізняється від 1, тим більше значимою буде модель.

5.6.3 Побудова моделі в системі Deductor

Імпортуйте файл із кредитними історіями в Deductor. На другому кроці Майстра імпорту у «Представление значений» укажіть Істина - True і Неправда – False.

Скоринг являє собою задачу бінарної класифікації, яка відносить позичальника до одному із двох класів — «поганий» або «гарний». Якщо позичальник «гарний» — кредит видається, якщо «поганий» — виноситься негативне рішення. Поділ позичальників на «поганих» і «гарних» здійснюється на основі якості обслуговування ними боргу, простіше говорячи — наявності прострочень. У банківській справі існують різні шкали переходу від числа прострочень до класу позичальника, і це тема для окремого обговорення. Приймемо наступне правило: якщо в клієнта було хоча б одне прострочення понад 60 днів, то його відносять до класу неблагонадійних. Запустіть Майстер обробки, у категорії Інші виберіть Калькулятор. Клацніть двічі на слові «Выражение» в колонці «Список выражений» і поміняйте параметри поля, що обчислюється, як зазначено на рис. 5.8.

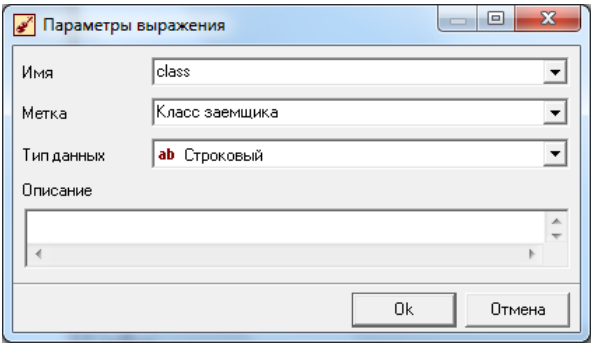


Рисунок 5.8 – Завдання параметрів поля, що обчислюється

У наступному стовпчику запишіть умову:

IF(COL12>0;"Поганий";"Гарний")

У результаті з'явиться нове поле, що обчислюється, — *Клас позичальника* (рис. 5.9).

Далі за допомогою візуалізатора Статистика можна довідатися, що є 500 записів з «поганими» кредитами, що становить 18,5 % усіх виданих кредитів. Це не так уже й мало: у практиці кредитного скоринга число записів міноритарного класу може бути й менше, аж до 1-3 %. Тому задача класифікації позичальників завжди вирішується в умовах сильної незбалансованості класів.

Таким чином, вихідна бінарна змінна — *Клас позичальника* — у нас уже є. У якості вхідних має сенс залишити всі, крім *Код* і *Дата:* очевидно, що вони ніяк не впливають на кредитоспроможність.

Поля *Вік* і *П/Д*, % залиште безперервними.

Побудуємо модель логістичної регресії, яка розрахує відповідні коефіцієнти регресії. Для цього викликаємо оброблювач Логістична регресія. Установіть вхідні й вихідні поля, як це показано на рис. 5.10.

У цьому ж вікні натисніть кнопку Настроювання нормалізації. Для вихідного поля *Клас позичальника* порядок сортування унікальних значень (яких у логістичній регресії завжди два) визначається типом події: перше — негативне, друге — позитивне. У скорингу прийняте, що чим вище рейтинг позичальника, тем вище кредитоспроможність, тому значення «гарний» буде позитивним результатом події (друге за рахунком), а «поганий» — негативним (перше).

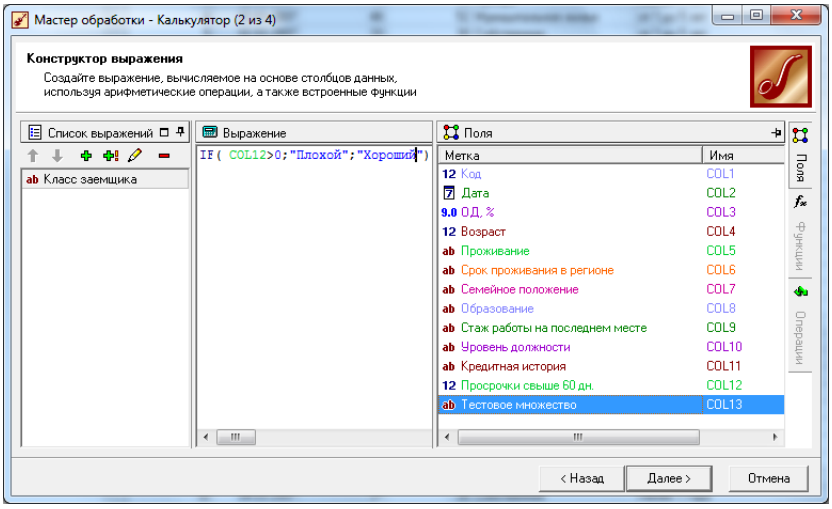


Рисунок 5.9 – Створення нового поля «Клас позичальника»

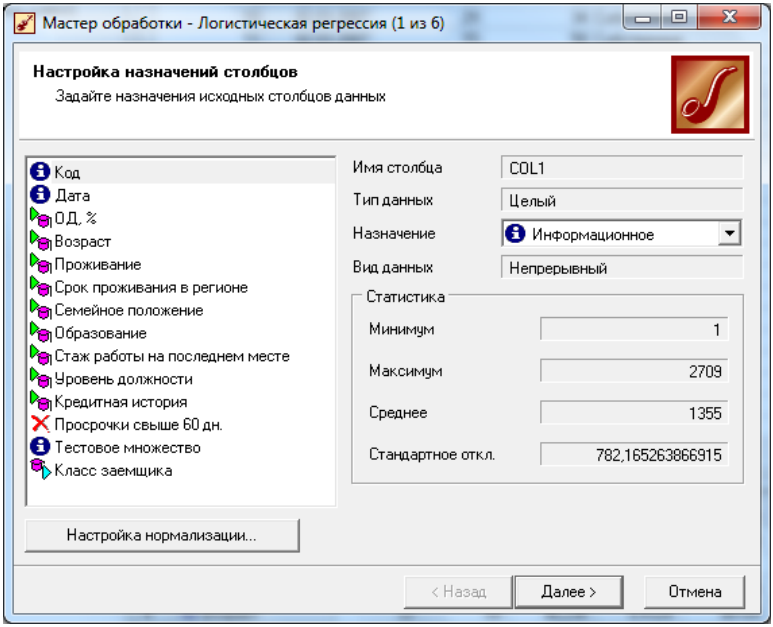


Рисунок 5.10 – Завдання вхідних і вихідних полів

Для вхідних стовпців (крім вимірів ПД% і Вік) укажіть спосіб кодування – ***комбінація бітів***.

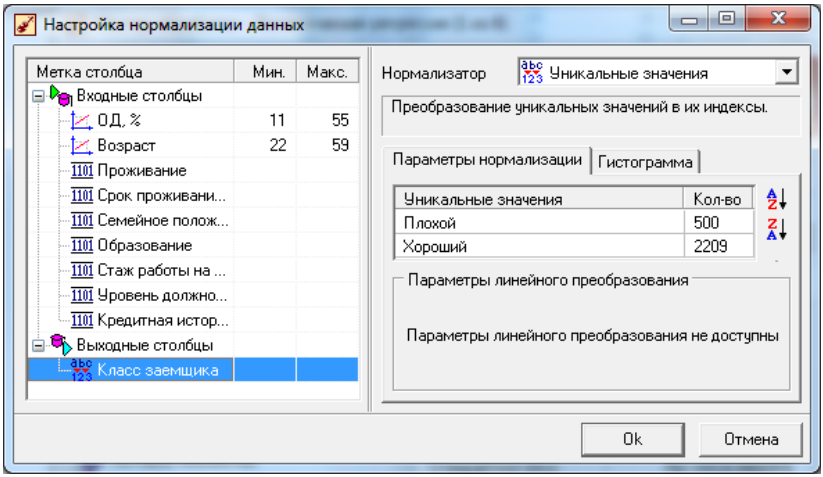


Рисунок 5.11 – Завдання типів подій вихідного поля

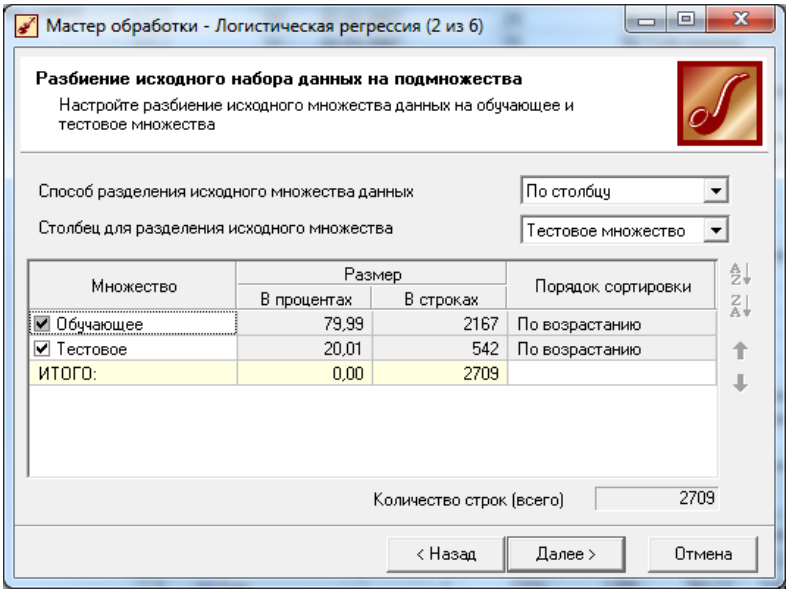


Рисунок 5.12 – Настроювання розбивки набору даних

У наступнім вікні майстра буде запропоновано настроїти навчальні й тестові множини. Оскільки в нас є спеціальне поле, у

якому зберігається інформація про розбивку на множини, укажемо його, установивши відповідні настроювання (рис. 5.12).

На третьому кроці майстра пропонується змінити параметри алгоритму логістичної регресії. За замовчуванням поріг класифікації дорівнює 0,5. Поки залиште всі параметри без змін.

На останньому кроці натисніть кнопку Пуск — буде побудована модель і майстер запропонує вибрати візуалізатори вузла. Укажіть наступні: ROC-анализ, Коэффициенты регрессии, Что-если, Таблица сопряженности, Таблица.

У візуалізаторі Таблиця видне, що додалися два нові стовпчики: Клас позичальника Рейтинг і Клас позичальника_OUT. Рейтинг являє собою розраховане значення y по рівнянню логістичної регресії, а друге поле визначає приналежність до того або іншого класу залежно від порога відсікання.

Візуалізатор Коефіцієнти регресії наочно показує розраховані коефіцієнти логістичної регресії, які є прототипом скорингової карти, відносини шансів, що їм відповідають (рис. 5.13).

Проінтерпретуємо відношення шансів для ознаки *Стаж роботи.* Якщо стаж роботи на останньому місці від 1 до 3 років, то шанси стати благонадійним позичальником при фіксованих значеннях інших змінних в OR = 1,71 рази вище в порівнянні з тем, у кого стаж менш 1 року. А якщо стаж понад 3 років, то шанси збільшуються в 2,67 рази.

Проінтерпретуємо тепер відношення шансів для поля *ОД, %: OR* = 0,96. Воно менше одиниці, виходить, збільшення кредитного навантаження знижує підсумковий скоринговий бал клієнта. Розглянемо потенційного позичальника А, у якого частка виплат по кредиту в структурі доходу на 10% менше, чим у позичальника Б.

Тоді можна сказати, що зниження щомісячних виплат на 10% приводить до того, що ймовірність стати гарним позичальником виростає в $e^{10 \cdot 0,041506} = 1,52$ рази.

Атрибут	Коэффициент	Отношение шансов
9.0 <Константа>	0,94342	
9.0 ОД, %	-0,041506	0,95934
12 Возраст	0,002304	1,0023
ab Проживание		
Аренда	0	1
Муниципальное жилье	1,5517	4,7196
Собственник	1,9372	6,9396
ab Срок проживания в регионе		
менее 1 года	0	1
от 1 до 5 лет	1,0084	2,7412
свыше 5 лет	1,4721	4,3583
ab Семейное положение		
Другое	0	1
Женат/замужем	1,2499	3,4901
Разведен(а)/Вдовство	-0,54479	0,57996
Холост/Не замужем	0,70513	2,0241
ab Образование		
высшее	0	1
среднее	-1,8317	0,16015
среднее специальное	-0,86975	0,41906
ab Стаж работы на последнем месте		
менее 1 года	0	1
от 1 года до 3 лет	0,54157	1,7187
свыше 3 лет	0,98386	2,6748
ab Уровень должности		
руководитель высшего звена	0	1
руководитель среднего звена	-0,067395	0,93483
сотрудник	-0,45748	0,63287
ab Кредитная история		
нет данных	0	1
отрицательная	-1,9173	0,147
положительная	2,052	7,7836

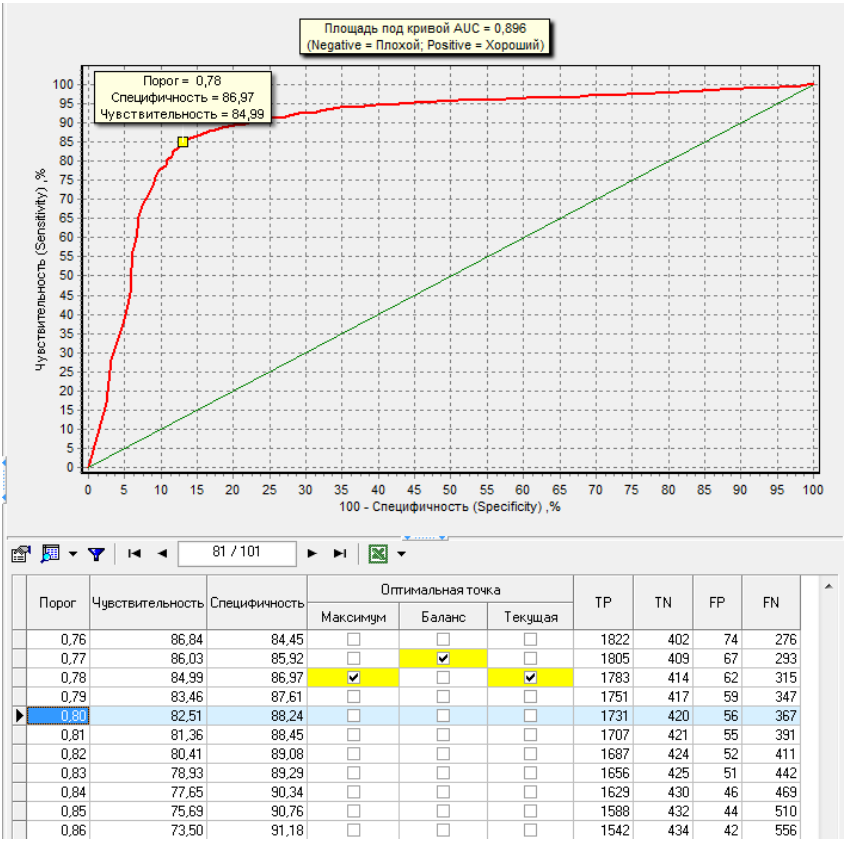
Рисунок 5.13 – Коефіцієнти логістичної регресії

Візуалізатор ROC-крива виводить графік ROC-кривої, на якім за замовчуванням відображаються положення поточного порога відсікання, а також значення чутливості й специфічності, показник AUC і типи подій (рис. 5.14). Площа під кривою дорівнює 0,894 на навчальній множині й 0,905 — на тестовій, що говорить про дуже гарну прогнозну здатність побудованої моделі.

Однак оптимальна точка для даної моделі не 0,5. Максимальна сумарна чутливість і специфічність досягається в точці 0,78 (для розрахунків і відображення оптимальної точки необхідно в меню кнопки `Тип оптимальної точки` вибрати пункт `Максимум`). Для установки нового порога відсікання, рівного 0,78, слід перенастроїти вузол-оброблювач логістичної регресії. У цій точці Se = 85 %, Sp = 86 %, що означає: 85 % благонадійних позичальників будуть виявлені класифікатором, а 100 — 86 = 14 % несумлінних позичальників одержать кредит. На тестовій множині спостерігається схожа картина: Se = 84 %, Sp = 86%.

У загальному випадку, проектуючи визначення чутливості й специфічності на скоринг (і враховуючи, що клас позичальника «гарний» відповідає позитивному результату), можна вважати, що скорингова модель із високою специфічністю відповідає *консервативній кредитній політиці* (частіше відбувається відмова у видачі кредиту), а з високою чутливістю — *політиці ризикованих кредитів.* У першому випадку мінімізується кредитний ризик, пов'язаний із втратами позички й відсотків і з додатковими витратами на повернення кредиту, а в другому — комерційний ризик, пов'язаний з упущеною вигодою.

Це добре ілюструє візуалізатор ***Таблиця спряженості*** (рис. 5.15), яка є не що інше, як матриця класифікації.



Порог	Чувствительность	Специфичность	Оптимальная точка			TP	TN	FP	FN
			Максимум	Баланс	Текущая				
0,76	86,84	84,45	☐	☐	☐	1822	402	74	276
0,77	86,03	85,92	☐	☑	☐	1805	409	67	293
0,78	84,99	86,97	☑	☐	☑	1783	414	62	315
0,79	83,46	87,61	☐	☐	☐	1751	417	59	347
0,80	82,51	88,24	☐	☐	☐	1731	420	56	367
0,81	81,36	88,45	☐	☐	☐	1707	421	55	391
0,82	80,41	89,08	☐	☐	☐	1687	424	52	411
0,83	78,93	89,29	☐	☐	☐	1656	425	51	442
0,84	77,65	90,34	☐	☐	☐	1629	430	46	469
0,85	75,69	90,76	☐	☐	☐	1588	432	44	510
0,86	73,50	91,18	☐	☐	☐	1542	434	42	556

Рисунок 5.14 – Графік ROC-кривої скорингової моделі

Класс заемщика

Фактически	Классифицировано		
	Плохой	Хороший	Итого
Плохой	345	55	400
Хороший	271	1496	1767
Итого	616	1551	2167

Рисунок 5.15 – Таблиця спряженості – робоча вибірка

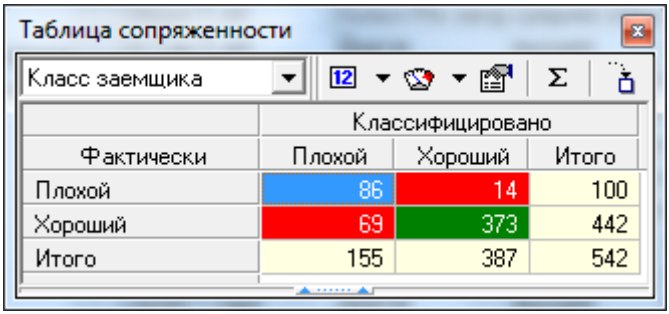


Рисунок 5.16 – Таблиця спряженості – тестова вибірка

З таблиці видно, що на навчальній множині модель частіше відмовляла у видачі кредиту «гарним» позичальникам, чим видавала кредит «поганим». Точність класифікації склала 85%. На тестовій множині спостерігається приблизно та ж картина (точність класифікації 84,7 %), а рівень схвалених кредитів (Approval Rate, *AR)* тут становить

$$AR = (387/542) \cdot 100\ \% = 72\ \%$$

при рівні дефолтних кредитів (Bad Rate, *BR*) рівному

$$BR = (14/387) \cdot 100\ \% = 3{,}62\ \%.$$

Якщо така ситуація не влаштовує, можна знизити поріг відсікання й добитися того, щоб модель частіше видавала позитивне рішення. Відсоток відмов поменшається, але зросте й кредитний ризик. Тому вибір точки відсікання залежить від поставлених цілей — знизити частку «поганих» кредитів або збільшити кредитний портфель, частіше виносячи позитивне рішення по клієнту.

Припустимо, нам відомі витрати помилкової класифікації: $C_{FN} / C_{FP} = 1/4$, тобто видача кредиту несумлінному позичальникові обходиться в 4 рази дорожче, чим відмова сумлінному. Тоді ми можемо, використовуючи правило Байєса, оцінити оптимальний скоринговий бал *P:*

$$P > 1/(1 + ¼) = 0{,}80.$$

Візуалізатор «Что-если» дозволяє побачити, як буде поводитися побудована модель при подачі на її вхід тих або інших даних. Інакше кажучи, проводиться експеримент, у якім, змінюючи значення вхідних нулів логістичної регресії, аналітик спостерігає за зміною значень на виході. Можливість аналізу за принципом «Что-если» особливо коштовна, оскільки дозволяє досліджувати правильність роботи системи, вірогідність отриманих результатів, а також її стійкість. Візуалізатор «Что-если» включає табличне й графічні уявлення, які формуються одночасно (рис. 5.17).

У верхній частині табличного уявлення відображаються вхідні поля, а в нижній — вихідні й розрахункові. Змінюючи значення вхідних полів, аналітик дає команду виконати розрахунки й спостерігає розраховані значення виходів логістичної регресії.

У графічному уявленні візуалізатора «Что-если» по горизонтальній осі діаграми відкладається весь діапазон значень поточного поля вибірки, а по вертикальній — значення відповідних виходів моделі. На діаграмі «Что-если» видно, при якім значенні входу змінюється значення на відповідному виході. Якщо, наприклад, у всім діапазоні вхідних значень вихідне значення для даного поля не змінювалося, то діаграма буде являти собою горизонтальну пряму лінію. У нашому випадку встановлена графічна залежність зміни кредитного рейтингу конкретного клієнта від коефіцієнта О/Д (усі інші входи — константи). Видно, що зі збільшенням О/Д рейтинг практично лінійно падає.

При бажанні від моделі логістичної регресії нескладно перейти до скорингової карти, для чого потрібно перевести коефіцієнти логістичної регресії в лінійну шкалу.

Підбираючи поріг відсікання, ми можемо встановити співвідношення рівня схвалень *AR* і очікуваної величини простроченої заборгованості *BR*.

Поле	Значение
Входные	
9.0 ОД, %	40
12 Возраст	26
ab Проживание	Аренда
ab Срок проживани...	от 1 до 5 лет
ab Семейное полож...	Другое
ab Образование	высшее
ab Стаж работы на ...	от 1 года до 3 лет
ab Уровень должно...	сотрудник
ab Кредитная истор...	нет данных
ab Тестовое множе...	False
Выходные	
ab Класс заемщика	Плохой
Расчетные	
9.0 Класс заемщика...	0,628227048047599

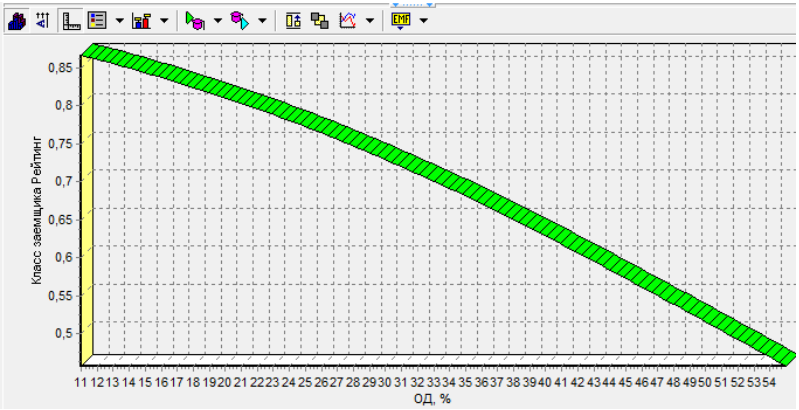


Рисунок 5.17 – Візуалізатор «Что-если»

5.7 Скорингова модель на основі дерева рішень

Тепер скористаємося іншим інструментом — деревом рішень. Використовуйте таблицю, у якій ви вже створили вихідне поле *Клас позичальника.*

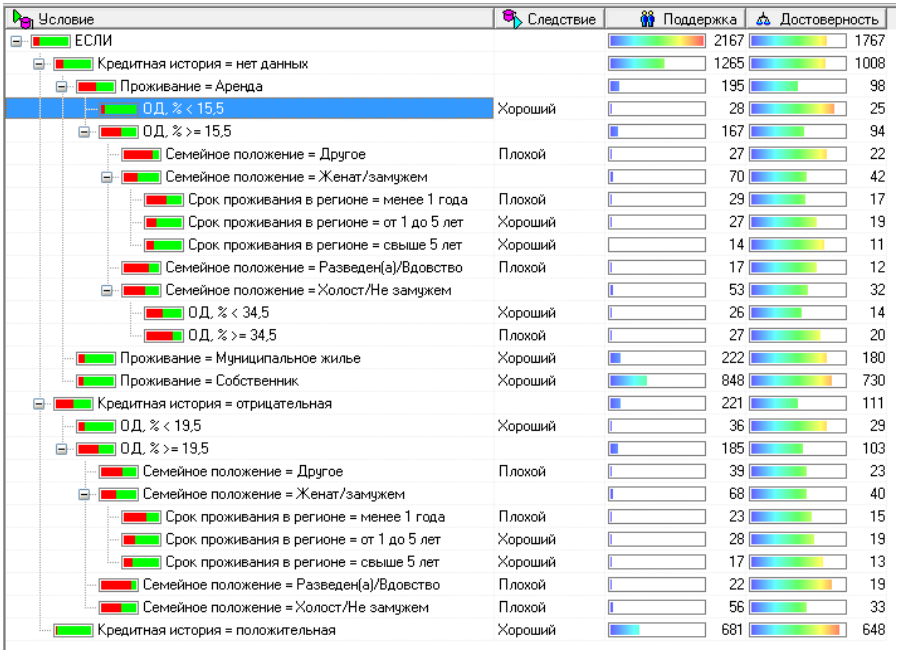
Додайте в сценарій однойменний вузол через *Майстер обробки.*

Наступні два кроки майстра аналогічні описаним раніше для вузла *Логістична регресія.* Відзначте поля Код і Дата як інформаційні, а поля Прострочення й Тестова множина – невикористовувані. На четвертому кроці відкриється вікно вибору параметрів алгоритму. Тут не міняйте настроювання, прийняті за замовчуванням, за винятком мінімальної кількості прикладів у вузлі, при якім буде створюватися новий. Задайте цей параметр рівним приблизно 1% від об'єму навчального множини (тобто 20); менше значення може привести до появи недостовірних правил, більше — до майже повної відсутності таких.

На наступному кроці в якості бажаного способу побудови дерева залиште режим автоматичної побудови. Запустивши його натисканням кнопки *Пуск,* пройдіть по кроках майстра далі й виберіть потрібні візуалізатори, відзначте прапорцями Дерево рішень, Значимість атрибутів, Що-Якщо, Таблиця спряженості.

У результаті роботи алгоритму було виявлено 18 правил. Точність класифікації на навчальній множині склала 85%, на тестовій — 87%. Візуалізатор *Дерево рішень* дозволяє побачити отриманий набір правил у схематичному виді, а також виводить показники вірогідності й підтримки для кожного вузла (рис. 5.18). Це і є скоринговая модель. Вона менш звична, оскільки тут не нараховуються бали за характеристики позичальника, але теж пояснює результат класифікації того або іншого позичальника.

У принципі, вірогідність кожного правила можна сприймати як підсумковий скоринговий бал з тим застереженням, що для поганих позичальників він дорівнює величині, отриманої вирахуванням з 100%-го значення вірогідності.



Условие	Следствие	Поддержка	Достоверность
ЕСЛИ		2167	1767
Кредитная история = нет данных		1265	1008
Проживание = Аренда		195	98
ОД, % < 15,5	Хороший	28	25
ОД, % >= 15,5		167	94
Семейное положение = Другое	Плохой	27	22
Семейное положение = Женат/замужем		70	42
Срок проживания в регионе = менее 1 года	Плохой	29	17
Срок проживания в регионе = от 1 до 5 лет	Хороший	27	19
Срок проживания в регионе = свыше 5 лет	Хороший	14	11
Семейное положение = Разведен(а)/Вдовство	Плохой	17	12
Семейное положение = Холост/Не замужем		53	32
ОД, % < 34,5	Хороший	26	14
ОД, % >= 34,5	Плохой	27	20
Проживание = Муниципальное жилье	Хороший	222	180
Проживание = Собственник	Хороший	848	730
Кредитная история = отрицательная		221	111
ОД, % < 19,5	Хороший	36	29
ОД, % >= 19,5		185	103
Семейное положение = Другое	Плохой	39	23
Семейное положение = Женат/замужем		68	40
Срок проживания в регионе = менее 1 года	Плохой	23	15
Срок проживания в регионе = от 1 до 5 лет	Хороший	28	19
Срок проживания в регионе = свыше 5 лет	Хороший	17	13
Семейное положение = Разведен(а)/Вдовство	Плохой	22	19
Семейное положение = Холост/Не замужем	Плохой	56	33
Кредитная история = положительная	Хороший	681	648

Рисунок 5.18 – Скоринговая модель – дерево рішень

Тепер відкрийте таблиці спряженості цього дерева рішень.

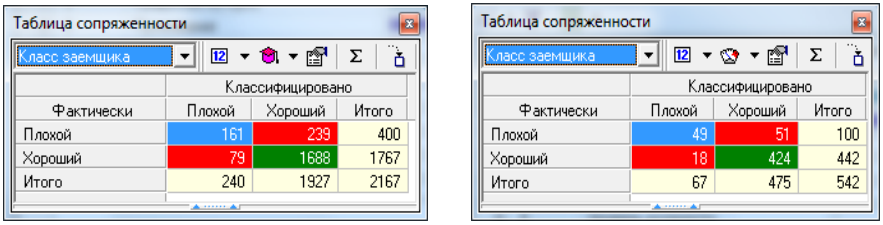


Таблица сопряженности

Класс заемщика

Фактически	Классифицировано: Плохой	Хороший	Итого
Плохой	161	239	400
Хороший	79	1688	1767
Итого	240	1927	2167

Таблица сопряженности

Класс заемщика

Фактически	Классифицировано: Плохой	Хороший	Итого
Плохой	49	51	100
Хороший	18	424	442
Итого	67	475	542

Рисунок 5.19 – Таблиці спряженості для робочої й тестової вибірок

Виявляється, у порівнянні з моделлю на основі логістичної регресії тут зовсім інша ситуація. Дерево рішень значно частіше схвалює неблагонадійних позичальників, тому що його побудова йде в умовах незбалансованості класів. У результаті частка дефолтних кредитів на тестовій множині дорівнює BR = 51/475 • 100 % = 10,7 %, що в 3 рази вище цього ж показника в моделі логістичної регресії

(правда, рівень схвалень виростає до 87,6%). Що робити, якщо така ситуація не влаштовує? У логістичній регресії для розв'язку цієї проблеми ми варіювали порогом відсікання, а в дереві рішень такої можливості немає.

Нам допоможуть спеціальні стратегії семплінга для зрівноважування навчальної множини: вибірка з дублюванням міноритарного класу (oversampling) і вибірка з видаленням прикладів мажоритарного класу (undersampling). Оскільки прикладів не так багато (400 — з поганими клієнтами й 1767 — з гарними) і інформація про кожного позичальника являє цінність, має сенс використовувати перший варіант — з дублюванням. Нехай відношення витрат помилкової класифікації залишиться колишнім: 1:4. Тоді, згідно із правилом, до навчальної вибірки потрібно додати 3•400 = 1200 прикладів, і загальне число записів складе 3367, а частка поганих збільшиться до 47 %.

Процедуру дублювання записів, що належать до міноритарного класу, потрібно здійснювати тільки на навчальній множині.

Для цієї операції знову залучимо кілька вузлів із групи Трансформація даних. Фільтр і Злиття даних (рис. 5.20).

Побудувавши дерево рішень по збалансованій вибірці, переконайтеся, що ситуація покращилася: тепер на тестовій множині модель частіше відмовляє у видачі гарним позичальникам, ніж схвалює поганих. Ці результати схожі з тими, які видає модель логістичної регресії.

Таким чином, ми одержали декілька скорингових моделей. Варіюючи порогами відсікання й застосовуючи спеціальні прийоми боротьби з незбалансованістю класів, можна підібрати ту модель, яка відповідає заданим потребам кредитної установи за рівнем схвалень заявок і очікуваній частці простроченої заборгованості. Перевіряти нових клієнтів можна за допомогою оброблювача Скрипт.

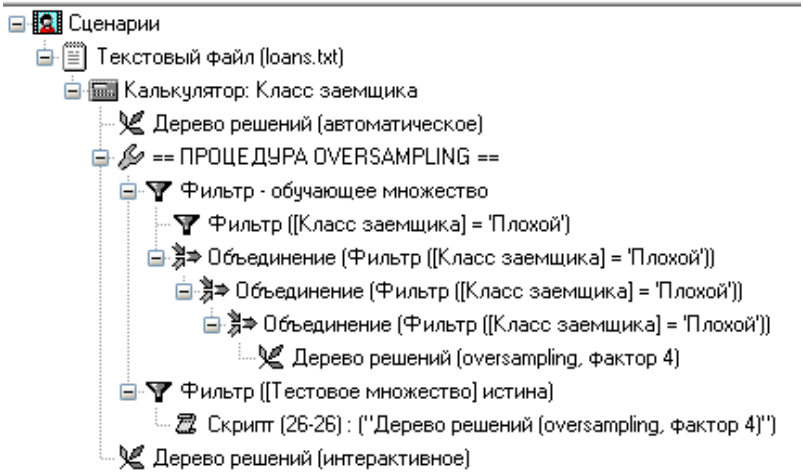


Рисунок 5.20 – Сценарій для побудови скорингової моделі на основі дерева рішень

Інтерактивне дерево рішень

До цього ми одержували дерево, яке будувалося автоматичним способом, тобто алгоритм на кожному кроці вибирав атрибут для розбивки за заданим критерієм. Відомо, що алгоритми побудови дерев «жадібні», тому не факт, що підсумкове дерево буде найкращим. У той же час іноді є експертні знання, які дозволяють «втрутитися» у процес формування дерева й вибору атрибутів, а також порогів для розбивки. Можливо, це й не підвищить точність моделі, але правила стануть більш логічними, з погляду експертів.

Кредитний скоринг являє собою той самий випадок, коли банківські аналітики мають певні знання й прагнуть, щоб у моделі розгалуження по атрибутах здійснювалося в певному порядку. Наприклад, якщо є атрибути *Наявність квартири* й *Вартість квартири*, то розумно відразу після першого розглянути другий. Ще приклад: після суми кредиту відразу бажане проаналізувати первісний внесок.

В аналітичній платформі Deductor є можливість побудови інтерактивних дерев рішень. Задамося метою побудувати скорингову

модель на попередній вибірці, прийнявши до уваги наступні побажання експертів.

1. Першим атрибутом, по якім аналізують позичальника, повинен бути атрибут *Кредитна історія.*

2. Далі необхідно розглянути коефіцієнт О/Д. Усіх клієнтів потрібно розбити на три категорії: позичальники з низьким О/Д (до 20 %), з помірним (від 20 до 40 %) і високим (від 40 %).

Додайте в сценарій новий вузол дерева рішень і на п'ятому кроці майстра поставте перемикач у позицію *Інтерактивний режим*.

У результаті візуалізатор ***Дерево рішень***, що відкрився, не буде містити жодного вузла. На панелі інструментів натисніть кнопку *Розбити поточний вузол на підвузли...,* відкриється відповідне вікно (рис. 5.21).

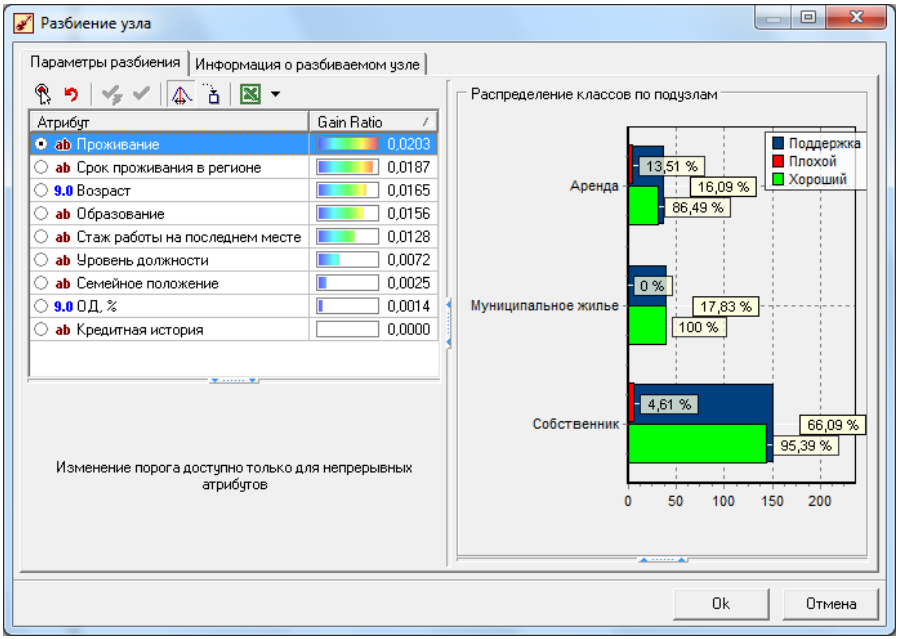


Рисунок 5.21 – Вікно вибору атрибута для розбивки в інтерактивному режимі: перший крок

Ліворуч у списку виводяться всі атрибути разом з розрахованими значеннями приросту інформації *Gain Ratio*, а праворуч — діаграми розподілу класів по підвузлам. За замовчуванням пропонується атрибут з максимальним значенням *Gain Ratio*, але його можна перевизначити. У цьому випадку нічого робити не потрібно, оскільки розбивка й так почнеться по атрибуту *Кредитна історія*. Натискання кнопки OK приведе до того, що в дерево додасться три вузли цього атрибута зі значеннями *немає даних, негативна, позитивна*.

Продовжимо розбивку далі, вибравши вузол

Кредитна історія = немає даних (рис. 5.22).

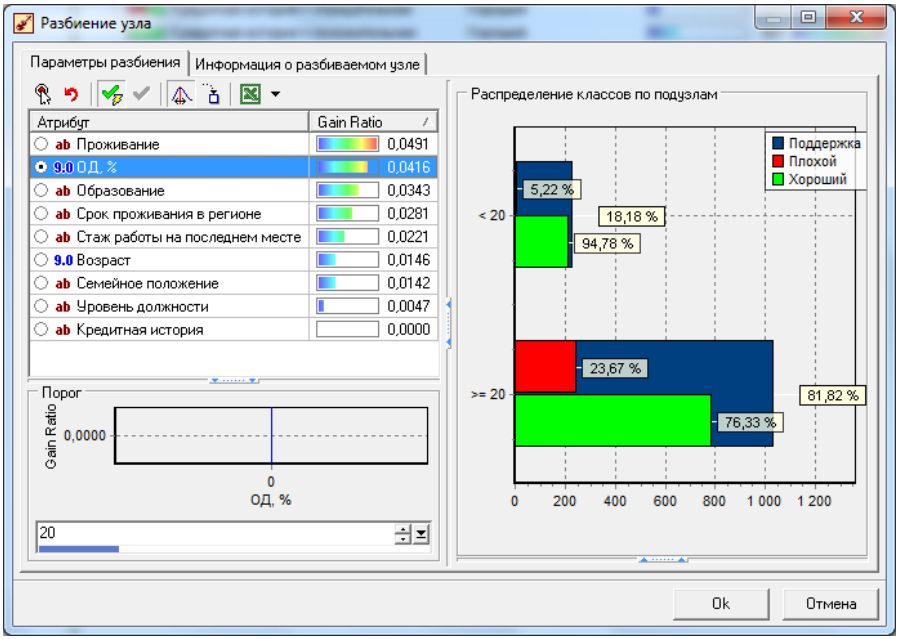


Рисунок 5.22 - – Вікно вибору атрибута для розбивки в інтерактивному режимі: другий крок

Тут у якості оптимального з погляду приросту інформації пропонується атрибут *Проживання*. Перевизначте його на *ОД, %,* указавши в нижній частині вікна поріг, рівний 20. Потім для вузла *ОД, % >20* знову виберіть розбивку по *ОД, %,* але вже з порогом

40, після чого натисніть кнопку *Побудувати дерево, починаючи з поточного вузла*. У результаті гілка дерева буде повністю готова (рис. 5.23).

Аналогічним чином добудовується дерево для вузлів, що залишилися. Якість класифікації, як і колись, можна оцінювати через таблиці спряженості.

Условие	Следствие	Поддержка	Достоверность
ЕСЛИ		2167	1767
Кредитная история = нет данных		1265	1008
О.Д, % < 20	Хороший	230	218
О.Д, % >= 20		1035	790
О.Д, % < 40		639	507
Проживание = Аренда		96	55
Срок проживания в регионе = менее 1 года	Плохой	44	32
Срок проживания в регионе = от 1 до 5 лет	Плохой	34	17
Срок проживания в регионе = свыше 5 лет	Хороший	18	12
Проживание = Муниципальное жилье	Хороший	111	90
Проживание = Собственник	Хороший	432	376
О.Д, % >= 40		396	283
Проживание = Аренда		62	37
Образование = высшее	Хороший	20	12
Образование = среднее	Плохой	18	12
Образование = среднее специальное	Плохой	24	17
Проживание = Муниципальное жилье	Хороший	70	49
Проживание = Собственник	Хороший	264	209
Кредитная история = отрицательная	Хороший	221	111
Кредитная история = положительная	Хороший	681	648

Рисунок 5.23 – Дерево рішень, побудоване в інтерактивному режимі

5.8 Завдання до лабораторної роботи

1. Використовуючи файл про кредитні історії, зібраний у банку - loans.txt, побудувати класифікатор на основі
   а) логістичної регресії;
   б) дерев рішень
   у програмі Deductor. Порівняти отримані моделі й оцінити їхню якість.

2. Побудувати моделі класифікації на даних, наведених нижче у варіантах завдань 5.1 - 5.9.

Варіант 5.1. Головним керівництвом економічного розвитку області був проведений вибірковий аналіз фінансового стану господарюючих суб'єктів, у результаті якого отримано три групи промислових підприємств: нормально функціонуючих, тих, що потребують фінансової підтримки й підприємства, які перебувають у стані банкрутства. Висновки щодо конкретного підприємства

робилися на основі аналізу коефіцієнта рентабельності (x_1), коефіцієнта поточної ліквідності (x_2), коефіцієнта забезпеченості власними коштами (x_3) і коефіцієнта втрати (відновлення) платоспроможності (x_4).

Таблиця 5.3 - ***Результати вибіркового аналізу фінансового стану господарюючих суб'єктів, що діють на території області***

Найменування підприємств	x_1	x_2	x_3	x_4
Група нормально функціонуючих підприємств				
Г.П. «Медтехніка»	8,09	1,30	0,23	1,13
Завод «Гама»	8,09	1,56	2,36	1,48
ВАТ «Іскра»	23,17	17,76	0,85	17,46
ВАТ «Автозапчастини»	2,10	28,78	0,97	31,02
ВАТ «Відеофон»	4,48	1,18	0,15	1,04
ЗАТ «Гідрогаз»	7,32	1,28	0,23	1,19
ЗАТ «Агропродукт»	12,00	1,89	0,47	1,79
ВАТ «Машобладнання»	4,45	7,52	0,87	7,42
Дорожні електромеханічні майстерні	2,79	2,00	0,50	1,69
ВАТ «Перетворювач»	1.32	10,02	0,24	9,46
Група підприємств, яким потрібна фінансова підтримка				
Завод «Радіоприлад»	0,52	0,95	-0,03	0,97
Виробничо-комерційна фірма «Флат»	2,84	0,98	-0,02	0,81
ВАТ «Суднoремонтний завод»	-84,86	2,02	0,50	1,99
ВАТ «Автодор»	34,8	9,82	-0,22	0,68
ВАТ «Рембутттехніка»	8,42	1,09	0,08	0,96
Група підприємств, які перебувають у стані банкрутства				

ВАТ «Втормет»	-2,13	0,73	-0,36	0,59
ВАТ «Вэлт»	-321,06	0,64	-1,02	0,72
ВАТ «ЗПП»	-48,53	0,97	-0,03	0,96
ВАТ «Тяжекс»	-356,24	0,32	-2,16	0,37
ВАТ «ЗСАК»	-41,47	0,92	-0,09	0,51

Потрібно, використовуючи наведені дані як навчальну вибірку, побудувати модель класифікації на основі одного з методів, розглянутих вище, а потім установити приналежність наступних підприємств до одного із трьох класів, визначивши тим самим його фінансовий стан.

Таблиця 5.4 - ***Показники фінансового стану господарюючих суб'єктів, які потрібно класифікувати***

Найменування підприємств	x_1	x_2	x_3	x_4
ЗАТ « КПП-Мікрон»	-5,17	2,97	-0,36	3,15
ВАТ Молочний комбінат	27,8	19,11	2,6	16,48
ТОВ Продовольча компанія	0,33	0,79	-0,61	0,51
ВАТ «Фруктові води»	-9,19	-0,1	0,19	0,51

Варіант 5.2. При оцінці ефективності діяльності підприємств легкої промисловості були отримано два класи підприємства: з високою й низькою продуктивністю праці. Крім того, були виявлені фактори, що визначають відповідний рівень продуктивності праці:

1) Частка робітників, зайнятих вручну не при машинах і механізмах у % (x_1);

2) Відсоток плинності кадрів (x_2);

3) Коефіцієнт змінності по всім робітникам (x_3);

4) Частка профільної продукції в загальному об'ємі продукції (x_4);

5) Електроозброєність, кВт (x_5).

В останньому стовпчику наведене модельне значення вироблення, тис.грн. (табл. 5.7). Опираючись на отримані результати:

1) проаналізуйте резерви росту продуктивності праці в групі гірших підприємств;

2) проведіть класифікацію підприємств, представлених у табл. 5.8.

Таблиця 5.5 - ***Класифікація підприємств за рівнем продуктивності праці***

Під-приєм-ство	Фактори					Значен. Вироб., тис.грн.
	x_1	x_2	x_3	x_4	x_5	
Група підприємств із високою продуктивністю праці						
1	34,1	11	1,47	93,4	21,3	69,75
2	33,7	12	1,29	91,7	32,2	63,89
3	23,6	23	1,17	95,3	27,8	60,64
4	29,6	12	1,47	95,3	22,6	81,62
5	25,3	16	1,44	96,3	21,9	81,02
6	17,9	27	1,52	91,1	27,8	62,53
7	29,3	13	1,62	94,5	23,5	83,55
Група підприємств із низькою продуктивністю праці						
8	38,4	12	1,36	96,4	15,5	61,68
9	37,5	15	1,44	95,2	12,3	54,86
10	32,2	19	1,29	96,1	16,4	54,71
11	26,1	19	1,46	92,3	11,4	55,90
12	30,7	25	1,57	92,9	23,4	51,34
13	28,7	18	1,47	85,2	22,7	41,14

Таблиця 5.6 - ***Характеристики підприємств, що підлягають класифікації***

Підприємство	Фактори				
	x_1	x_2	x_3	x_4	x_5
14	31,5	19	1,76	92,7	18,4
15	31,2	23	1,37	94,6	17,9
16	19,7	20	1,52	96,2	27,1
17	28,7	24	1,56	92,2	27,8

Варіант 5.3. Хворі гіпертиреозом (збільшення щитовидної залози) загальним числом 23 людини були розділені на три групи:

- *Група 1*. Лікування виявилося успішним; проведене через великий проміжок часу клінічне обстеження показало, що пацієнт здоровий.
- *Група 2*. Лікування безуспішне, тобто стан хворого залишився без зміни.
- *Група 3*. Результат лікування успішний, але надалі можливий рецидив.

За результатами обстеження 23 пацієнтів є наступні виміри:

- ***y6*** йод, що регіструється через 3 години після прийняття іспитової дози;
- ***y9*** йод, що регіструється через 48 годин після прийняття іспитової дози;
- ***y10*** зміст у крові білкове пов'язаного йоду (РВ - ^{131J}) через 48 годин.

Таблиця 5.7 – ***Результати обстеження пацієнтів***

N	**Гр.**	**y6**	**y9**	**y10**	**N**	**Гр.**	**y6**	**y9**	**y10**
1	1	14.4	25.1	0.20	12	1	47.5	50.1	0.29
2	1	20.1	40.1	0.11	13	1	54.0	57.0	0.19
3	1	24.1	32.1	0.17	14	1	16.1	20.6	0.22
4	1	11.1	16.9	0.12	15	1	57.5	74.5	0.49
5	1	16.3	32.1	0.36	16	1	37.8	63.0	0.32

6	1	40.5	64.4	0.21	17	2	55.8	48.0	2.74
7	1	52.7	50.0	0.53	18	2	75.0	60.0	1.37
8	1	20.8	22.3	0.13	19	2	72.0	65.0	0.70
9	1	14.0	3.1	0.18	20	2	70.6	45.0	1.40
10	1	27.0	41.7	0.19	21	3	24.1	45.0	0.22
11	1	44.3	63.8	0.22	22	3	33.2	55.0	0.01
					23	3	30.4	44.6	0.09

Побудувати модель класифікації для нових пацієнтів за результатами обстеження.

Варіант 5.4. Прикладом буде діагностика діабету (набір даних узятий з UCI machine learning repository). Навчальна вибірка містить 768 записів з наступними полями:

1. Число випадків вагітності;
2. Концентрація глюкози;
3. Артеріальний діастоличний тиск, мм. рт. ст.;
4. Товщина шкірної складки триглавого м'яза, мм.;
5. 2-х вартовий сироватковий інсулін;
6. Індекс маси тіла;
7. Числовий параметр спадковості діабету;
8. Вік, років;
9. Залежна змінна (1 – наявність захворювання, 0 – відсутність).

Розподіл залежної змінної такий: 500 випадків відсутності захворювання, 268 – його наявність.

Дані знаходяться в файлі diabet.txt.

Побудувати модель класифікації для нових пацієнтів за результатами обстеження.

Варіант 5.5. Побудувати модель класифікації крабів за статтю. Відома вибірка з 6 характеристиками краба та відомою статтю, що знаходиться в файлі crabdata.csv. В файлі маємо такі фізичні характеристики крабів:

- - різновид краба, розглядається тільки 2 різновида;
- - розмір передної губи краба;
- - ширина заднього панциря краба;
- - довжина краба;

- - ширина краба;
- - висота краба.

Останній стовпчик вказує стать краба.

Варіант 5.6. За моделлю Тафлера фінансова стійкість підприємства може бути оцінена за такими показниками:

- X1 - прибуток до виплат / поточні зобов'язання;
- X2 - поточні активи / зобов'язання;
- X3 - поточні зобов'язання / загальна вартість активів;
- X4 – інтервал кредитування.

У файлі «tafler.txt» перебувають зазначені дані для 103 підприємств. Побудувати модель класифікації та оцінити її якість.

Варіант 5.7. Згідно з моделлю Ліса фінансова стабільність підприємства може бути оцінена за такими показниками:

- X1 - обіговий капітал / сума активів;
- X2 - прибуток від реалізації / сума активів;
- X3 - нерозподілений прибуток / сума активів;
- X4 - власний капітал / позиковий капітал.

У файлі «lisy.txt» перебувають зазначені дані для 103 підприємств. Побудувати модель класифікації та оцінити її якість.

Варіант 5.8. Це завдання взято з UCI MLR (UCI Machine Learning Repository) - це відкрита колекція реальних наборов даних для перевірки і тестування різних алгоритмів машинного навчання. Усі набори даних поділені по категоріям (класифікація, кластеризація і ін.) та областям використання (наука, медицина, бізнес і ін.).

Веб-адреса репозиторія: http://archive.ics.uci.edu/ml/

Приклад використовує дані Congressional Voting Records Data Set (1987) - на підставі результатів 16 голосувань необхідно передбачити політичну приналежність американських сенаторів (республіканець або демократ). Дані знаходяться в файлі «Голосование конгресса.txt»

Відкривши статистику, побачимо, що всього в наборі 435 записів, з них 267 - демократи й 168 - республіканці.

Таблиця містить наступні поля : **Клас** – клас голосуючого (демократ або республіканець), інші поля інформують про те, як

голосували сенатори за прийняття різних законопроектів (так, ні, утримався).

Основною метою аналітика є віднесення сенатора до тієї або іншої партії. Механізм віднесення повинен бути таким, щоб сенатор указав, як він буде голосувати за різні законопроекти, а класифікатор відповістить на запитання, хто він – демократ або республіканець.

Варіант 5.9. Власникам компанії ВАТ «Спектр» належить мережа супермаркетів. В 20ХХ році ця компанія здійснювала торговельну діяльність на території 12 регіонів країни. У стратегічні плани компанії наступного року входить розширення мережі супермаркетів за рахунок освоєння нових ринків збуту в інших регіонах. Аналітиками компанії були ідентифіковані найбільш значимі для розв'язуваної задачі показники, що характеризують соціально-економічний розвиток регіонів. Такими показниками виявилися:

1) загальний обсяг товарообігу й платних послуг на душу населення (тис. грн.), x_1;

2) об'єм інвестицій в основний капітал на душу населення (тис. грн.), x_2;

3) коефіцієнт щільності автомобільних доріг, x_3;

4) співвідношення середньодушових доходів і середньодушового прожиткового мінімуму, x_4.

Беручи до уваги той факт, що в деяких регіонах компанія мала позитивний (довгочасне одержання прибутку) або негативний (зазнавала збитків) досвід своєї діяльності, ці регіони були розділені, відповідно, на дві групи. У результаті була сформована таблиця 5.8.

Показники, що характеризують рівень соціально-економічного розвитку регіонів, що залишилися, на території яких ВАТ «Спектр» ще не здійснював свою діяльність, але які входять у коло його комерційних інтересів, представлені в таблиці 5.9.

Фактично завдання полягає в одержанні прогнозних оцінок у номінальній шкалі, які дозволили б без проведення повномасштабних польових маркетингових досліджень передбачити успішність діяльності компанії в регіонах, зазначених у табл. 5.9.

Таблиця 5.8 - ***Показники, що характеризують рівень соціально-економічного розвитку регіонів в 20XX р.***

№	Регіон	x1	x2	x3	x4
Група регіонів, у яких діяльність ВАТ "Спектр" була успішною					
1.	Луганська область	28,94	8,64	32,06	2,29
2.	Донецька область	31,59	3,96	25,56	2,16
3.	Кіровоградська область	23,63	6,33	30,05	1,79
4.	Київська область	23,62	8,22	29,69	1,62
5.	Херсонська область	21,43	5,78	27,57	1,59
6.	Одеська область	17,62	4,62	24,62	1,57
7.	м. Київ	86,02	20,37	61,69	5,09
Група регіонів, у яких діяльність ВАТ "Спектр" не була успішною					
1.	Хмельницька область	17,97	2,45	28,41	1,41
2.	Вінницька область	14,07	3,94	25,86	1,22
3.	Миколаївська область	11,33	2,06	21,73	0,84
4.	Івано - Франківська область	15,93	4,76	31,05	1,31
5.	Тернопільська область	20,18	2,8	25,92	1,53

Таблиця 5.9 - ***Показники, що характеризують рівень соціально-економічного розвитку регіонів, щодо яких необхідно прийняти маркетингове рішення***

№	Регіон	x1	x2	x3	x4
1.	Чернівецька область	17,47	5,97	28,17	1,29
2.	Львівська область	14,88	6,28	15,78	1,32
3.	Рівненська область	16,27	7,8	29,91	1,32
4.	Житомирська область	23,16	8,2	37,83	1,62
5.	Ужгородська область	15,39	6,82	41,28	1,11
6.	Черкаська область	19,28	9,68	27,79	1,82

5.9 Контрольні питання

1. Що таке задача класифікації? Приведіть приклади з економіки, де виникають задачі класифікації.
2. Які дані потрібно мати для побудови моделі класифікації?
3. Як оцінюють якість алгоритмів класифікації?
4. У чому полягає задача кредитного скорінга?
5. Запишіть математичну модель логістичної регресії. Які задачі можна вирішувати на основі цієї моделі?
6. Як кодують категоріальні змінні при розв'язку задачі класифікації?
7. Що таке дерево рішень? Які алгоритми побудови дерев рішень ви знаєте?
8. На підставі яких критеріїв вибирають змінну для розгалуження при побудові дерев рішень?
9. Які методи скорочення дерева рішень і зупинки побудови дерева ви знаєте?
10. Навіщо будують дерево в інтерактивному режимі?

6 РЕКОМЕНДОВАНА ЛІТЕРАТУРА

1. Чубукова И.А. Data mining: учебное пособие – М.: Интернет-университет информационных технологий: БИНОМ: Лаборатория знаний, 2006. – 382 с. – ISBN 5-9556-0064-7.
2. Ситник В.Ф. Інтелектуальний аналіз даних. К.: КНЕУ, 2007. –
3. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – СПб.: БХВ-Петербург, 2007. – 384 с.
4. Паклин Н.Б., Орешков В.И. Бизнес-аналитика: от данных к знаниям: Учеб. пособие. 2-е изд., перераб. и доп. - СПб.: Питер, 2010. – 704 с.