НАВЧАЛЬНО-НАУКОВИЙ КОМПЛЕКС

«ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ»

НАЦІОНАЛЬНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

ПРОЕКТ
з курсу «Аналіз фінансово-економічних даних»

Виконали:

студенти 4 курсу,

групи КА-83

Костенко М.О.

Байбара А. Г.

Мартинченко А.О.

Прийняла:
Кузнецова Н.В,

Київ – 2021

Мета роботи: навчитися застосовувати методи інтелектуального аналізу даних для вирішення задач аналізу фінансово-економічних даних.

Використати апарат мереж Байєса для моделювання задач імовірнісного характеру і прогнозування ймовірнісних змінних, а також дерева рішень, регресійні моделі для прогнозування абсолютних значень фінансових характеристик та порівняти їх для обраного набору даних.

Порядок виконання роботи

Проект складається з обов'язкової та частини роботи за варіантом (яка виконується групою студентів у команді згідно заданого викладачем варіанту та набору вхідниих даних).

Обов'язкова частина роботи полягає у вирішенні сукупності конкретних задач та завантаженні їх на Гугл-клас (протоколу та програми). Частина роботи за варіантом передбачає використання декількох методів інтелектуального аналізу даних для прогнозування фінансово-економічних даних (на кожного студента по 1му методу) згідно встановлених етапів.

Командна частина роботи згідно отриманого варіанту

Етап 1. Використання мереж Байєса для моделювання конкретних

прикладних задач.

Мета етапу: навчитися застосовувати апарат мереж Байєса для

моделювання задач імовірнісного характеру.

Порядок виконання роботи

1. Власна постановка задачі моделювання.

➢ Виконати аналіз вхідних даних та встановити наявність

причинно-наслідкового зв'язку між компонентами даних.

2. Розбити навчальні дані на навчальну та перевірочну вибірку. В

якості навчальної вибірки використати 90% навчальних даних, а для

перевірки якості моделі використати 10% навчальної вибірки.

3. За навчальною вибіркою побудувати початкову структуру мережі

Байєса.

4. Провести навчання структури і параметрів мережі Байєса.

5. Сформувати висновок – визначити ймовірність події, яка заданапостановці задачі моделювання.

6. Обчислити загальну похибку моделі та похибки класифікації.

7. Ітеративно перерозбити вхідний набір на навчальну та перевірочну вибірку у різному співвідношенні. Виконати пп. 3-6 на нових вибірках. Чи вдалось покращити результати прогнозування? На скільки? Яке оптимальне розбиття? Встановити оптимальне співвідношення начальної та перевірочної вибірки для отримання найвищої загальної точності та/або помилок 1-го та 2-го роду або вищої якості моделі.

Етап 2. Використання методу дерев рішень для аналізу фінансово-

економічних даних та прогнозування фінансових показників

Мета етапу: навчитися будувати скорингові моделі на основі дерев

рішень

Порядок виконання роботи

1. Для набору даних згідно варіанту команди побудувати скорингову

модель у вигляді дерева рішень.

При цьому 90% даних використати для побудови, а 10% для

перевірки прогнозуючих якостей моделі.

В якості значення порогу при класифікації розгляньте випадки 95%,

90%, 85 та 80%.

4. Обчисліть загальну похибку моделі (CA – common accuracy).

5. Обчислити похибки класифікації – 1-го, 2-го роду та загальну.

6. Спрогнозуйте для перевірочної вибірки значення та обчисліть

точність і якість прогнозу.

Етап 3. Застосування регресійних моделей для аналізу та

прогнозування фінансових показників

1. Побудувати набір регресійних моделей для прогнозування фінансових даних на основі навчальної вибірки і обрати кращу з них.

2. За перевірочною вибіркою оцінити якість моделей.

3. Оцінити параметри моделі, застосувати різні методи включення

факторів, покращити якість побудованих моделей.

4. Записати рівняння регресійної моделі на основі оцінених параметрів

та вхідних змінних.

5. Застосувати регресійну модель для прогнозування фінансових

показників на перевірочній вибірці.

Етап 4. Порівняння різних методів і підходів для аналізу фінансово-

економічних даних

1. Порівняти всі побудовані на етапах 1-3 моделі та зробити

висновки щодо доцільності використання застосованих моделей для

прогнозування фінансово-економічних даних та в яких випадках, які моделі доцільно застосовувати. В разі, якщо у команді 5 осіб, то додатково застосувати нейронні мережі (різних типів) для прогнозування фінансових показників за схемою етапу 3 пп.1–5.

2. Підготувати фінальний звіт з обов'язковою та командною

частиною та задокументованими етапами 1–4.

3. Завантажити звіт та програми з наборами даних у Google

Classroom.

4. Підготувати презентацію Вашого проекту (командної роботи) та представити її на парі у вигляді публічного захисту.

Командне завдання

Встановити причини відтіку клієнтів банку. Надані данні:
RowNumber - відповідає номеру запису (рядка) і не впливає на вихідні дані.
CustomerId - містить випадкові значення і не впливає на вихід клієнта з банку.
Surname - прізвище клієнта не впливає на його рішення залишити банк.
CreditScore - може вплинути на відтік клієнтів, оскільки клієнт з вищим кредитним балом рідше виходить з банку.
Geography - місцезнаходження клієнта може вплинути на його рішення залишити банк.
Gender - цікаво дослідити, чи відіграє роль стать, на відтік з банку.
Age - це, безумовно, актуально, оскільки старші клієнти рідше залишають свій банк, ніж молодші.

Tenure - означає кількість років, протягом яких клієнт був клієнтом банку. Як правило, клієнти старшого віку більш лояльні та рідше йдуть з банку.
Balance - також дуже хороший показник відтоку клієнтів, оскільки люди з вищим балансом на рахунках рідше залишають банк у порівнянні з тими, хто має нижчий баланс.
NumOfProducts - стосується кількості продуктів, які клієнт придбав через банк.
HasCrCard - означає, чи є у клієнта кредитна картка. Ця графа також актуальна, оскільки люди з кредитною карткою рідше залишають банк.
IsActiveMember - активні клієнти рідше залишають банк.
EstimatedSalary - як і на балансі, люди з нижчими зарплатами частіше залишають банк у порівнянні з тими, у кого вища зарплата.
Exited - незалежно від того, вийшов клієнт з банку чи ні.

Банку потрібно знати, що веде клієнта до рішення залишити компанію. Запобігання відтоку дозволяє компаніям розробляти програми лояльності та кампанії утримання, щоб утримати якомога більше клієнтів.

Попередня обробка данних

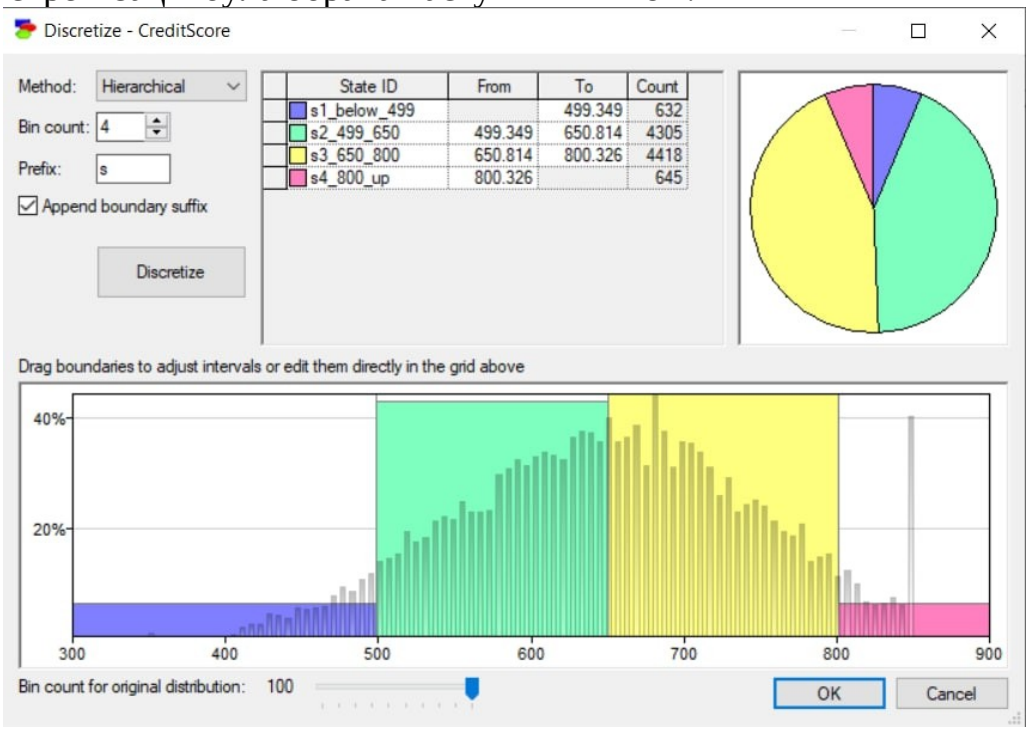
Оскільки банку потрібно знати, що веде клієнта до рішення залишити компанію, ми будемо розглядати вплив інших змінних на значення Exited, яке за данних умов є ключовим.
Одразу потрібно зазначити, що серед отриманних данних не усі впливають на результат, так, наприклад, значення колонок RowNumber, CustomerId, Surname які позначають номер запису, ID клієнта та його фамілію не впливають на його рішення залишити банк. Отже, при побудуванні мережі одразу потрібно заборонити наслідкові відношення між ними.
З іншого боку, показники балансу рахунку, активності клієнта, наявніть кредитної картки чи показник кредитного балу мають прямий вплив на рішення клієнта, і їх обовязково потрібно включити в подальший аналіз.
Серед наданих даних немає пропущенних даних, перегляд графічного представлення також показав, що серед данних немає аномально великих чи низьких значенн.
Отриманні дані складають 10000 записів, серед яких 2037 клієнтів відмовились від послуг банку, а інші 7963 продовжили утримувати рахунок. Тобто, відсоткове співвідношення між клієнтами, що залишились, і тими, хто закрив рахунок становить 20,27% \ 79.63%.
Початкову вибірку вирішено розбити на навчальну та тестову у співвідношенні 9:1, тобто 9000 записів у навчальній та 1000 у тестовій. Також важливо зберегти відношення закритих рахунків, тобто навчальна вибірка має складатися із 7119 збережених рахунків та 1881 закритих, а тестова — 842 відкритих та 158 закритих рахунків відповідно.

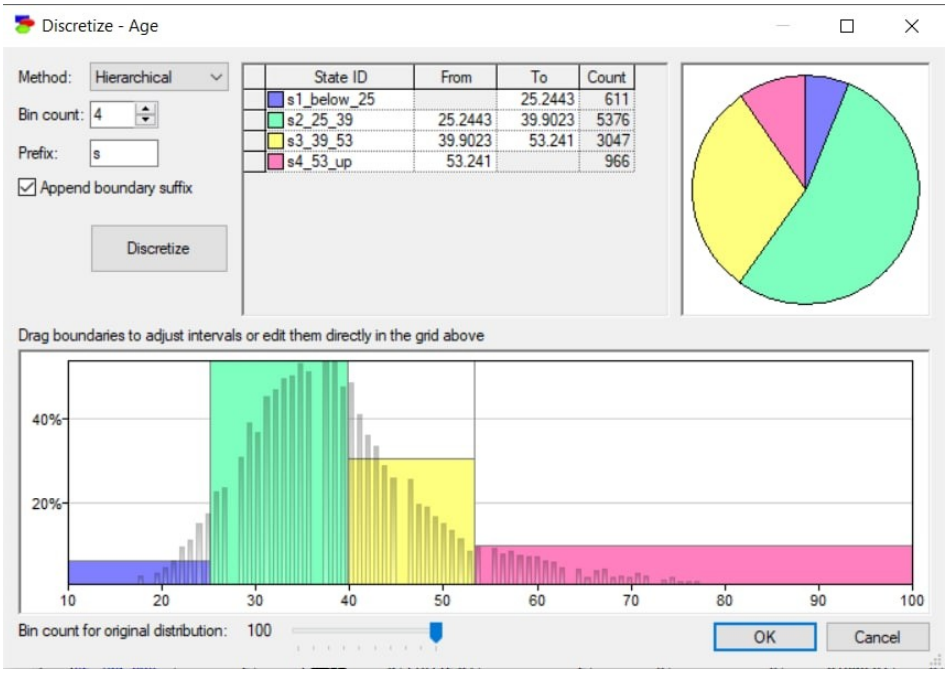
Також необхідним етапом буде зміна строкових змінних на числові для зручності обчислень та якості обробки. Отже, обрані зміни виглядають наступним чином:

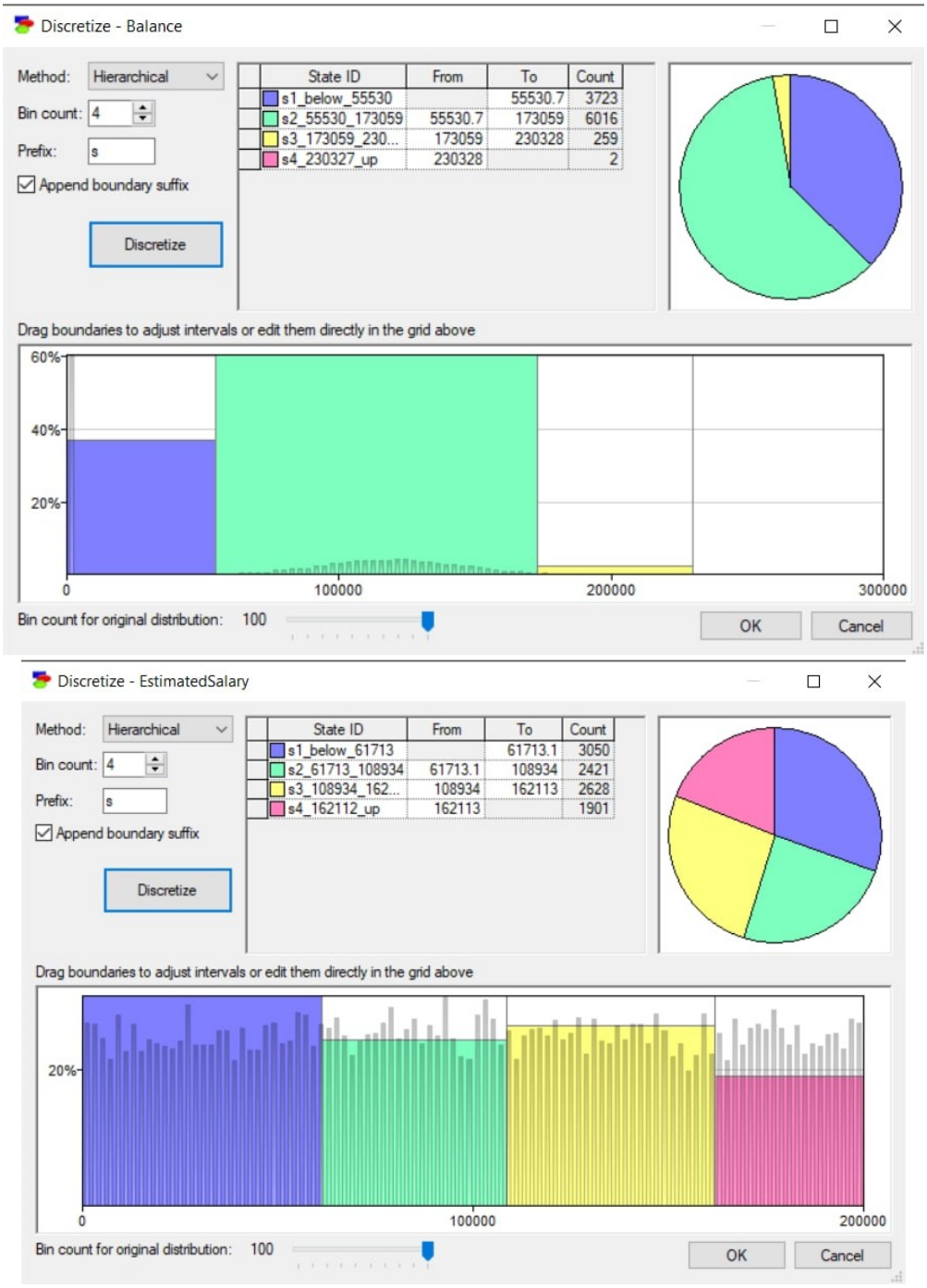
Female – 0
Male – 1

France - 1
Germany — 2
Spain — 3

Данні також необхідно дискретизувати для їх коректного використання у мережах Байєса. Дискретизація була обрана наступним чином:







Використанні методи

За умовою роботи необхідно було використати у аналізі дерева рішень та регрессійну модель. Для аналізу змінної Exited було використано логістичну регресію (англ. logistic regression), оскільки цей статистичний регресійний метод застосовують у випадку, коли залежна змінна є бінарною, тобто може набувати тільки двох значень (0 або 1). Також, При запровадженні порогового значення може знаходити застосування у класифікуванні, і ця особливість була використана при фнальній побудові результатів.
Наведемо огляд використаних методів:

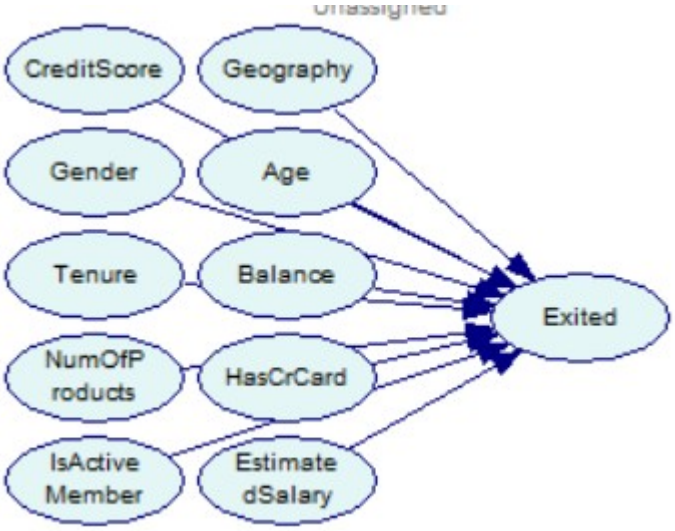
Мережі Байєса:
- Bayesian search
- Greedy ThickThinning
- PC

Дерева:
- CHAID
- Exhaustive CHAID
- CRT
- QUEST

Логістична регресія:
- Enter
- Forward Ward
- Forward Conditional
- Forward LR
- Backward Ward
- Backward Conditional
- Backward LR

Практичне завдання — Мережі Байєса

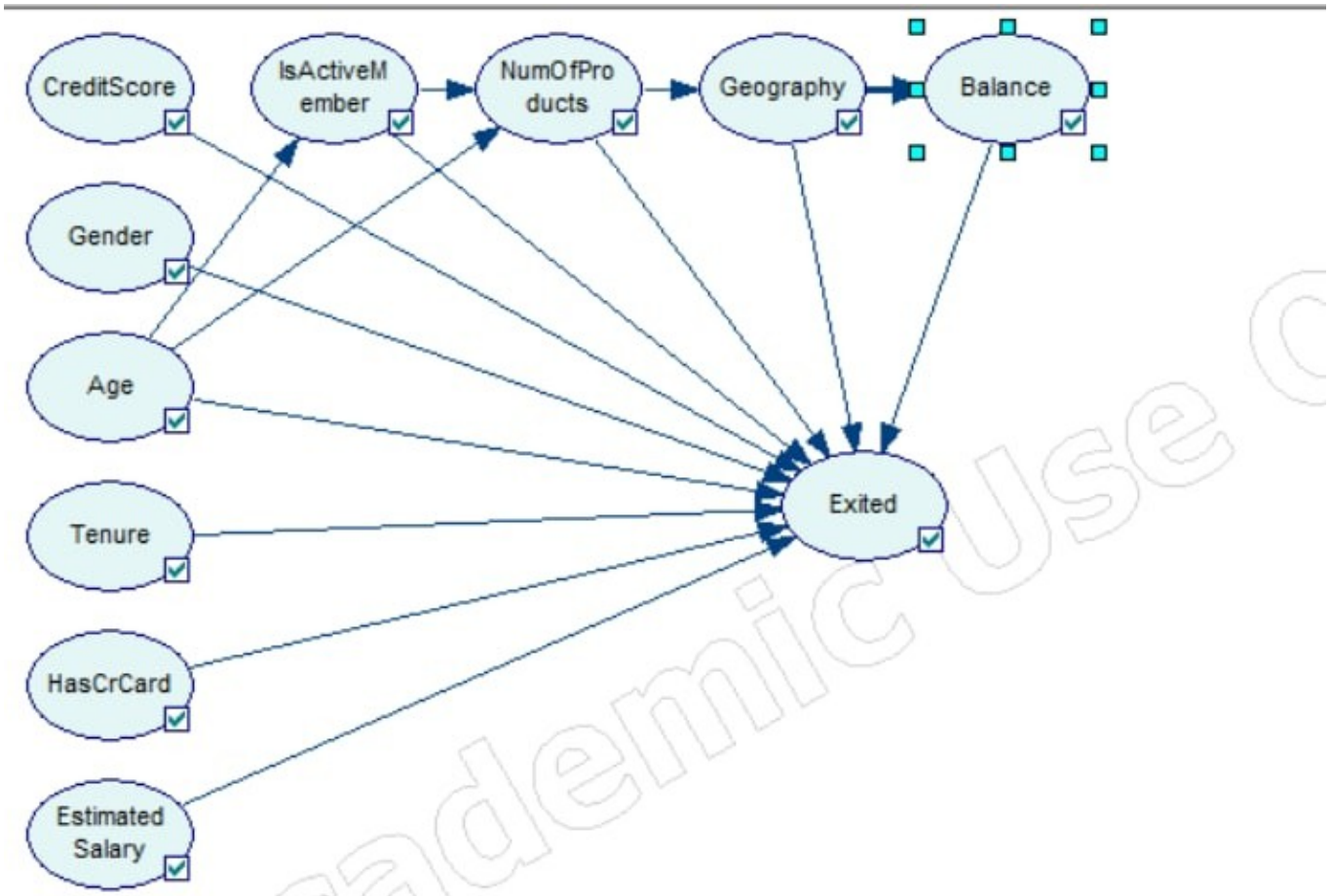
Мережі були побудовані із використанням трьох різних алгоритмів. Враховуючи надані данні, у схему також було внесено експертне знання у вигляді співвідношенн між змінними.



Розглянемо отриманні у побудові дерева та результати їх прогнозів:

Greedy ThinkThinning
Алгоритм навчання структурі Greedy Thick Thinning (GTT) заснований на байєсіанському підході пошуку. GTT починається з порожнього графіка і багаторазово додає дугу (без створення циклу), що максимально збільшує граничну

ймовірність P(D|S), поки жодне додавання дуги не призведе до позитивного збільшення (це фаза потовщення). Потім він багаторазово видаляє дуги, поки видалення дуги не призведе до позитивного збільшення P(D|S) (це фаза витончення). Це приблизний, але дуже швидкий алгоритм, який дає досить хороші структури. Ось побудоване дерево для алгоритму Greedy Thick Thinning:

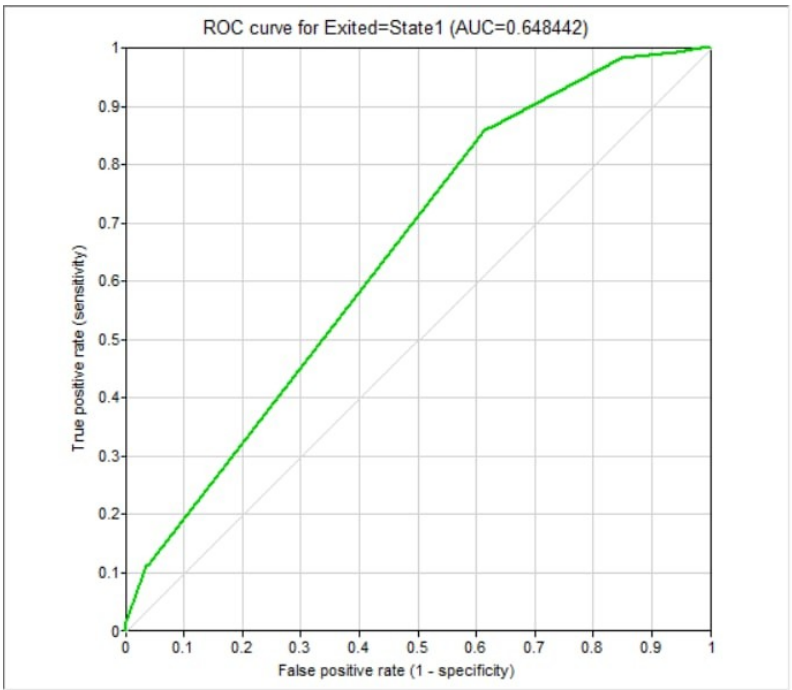


Input file: data_discret_learn.csv
Data rows: 8998
Elapsed time: 0.156s

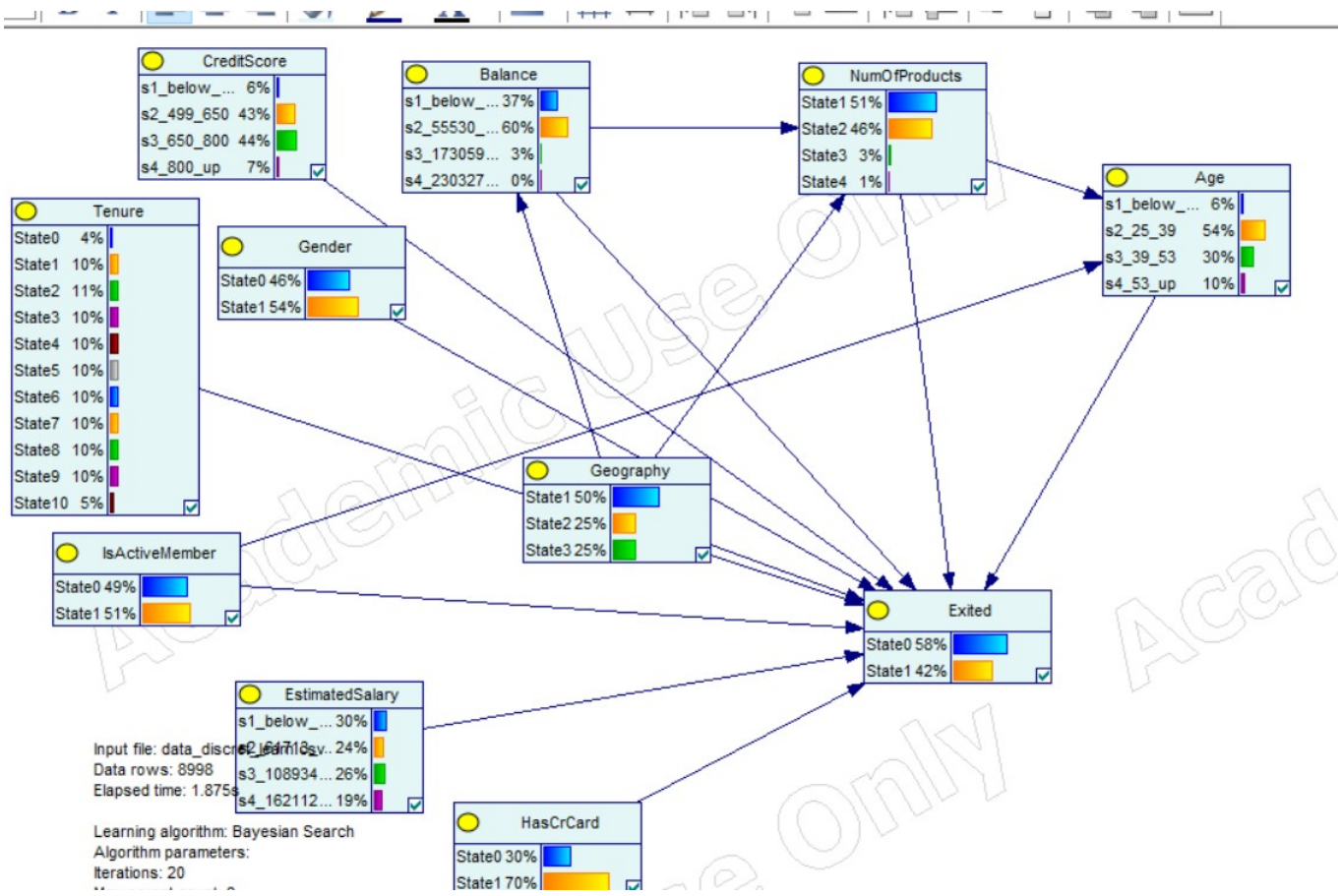
Learning algorithm: Greedy ThickThinning
Algorithm parameters:
Max parent count: 8
Background knowledge was provided:
forced arcs: 10

Score: -101027
EM Log Likelihood: -95209.8

Результати валідації:
Exited = 0.810379 (812/1002)
State0 = 0.962333 (792/823)
State1 = 0.111732 (20/179)
ROC curve:



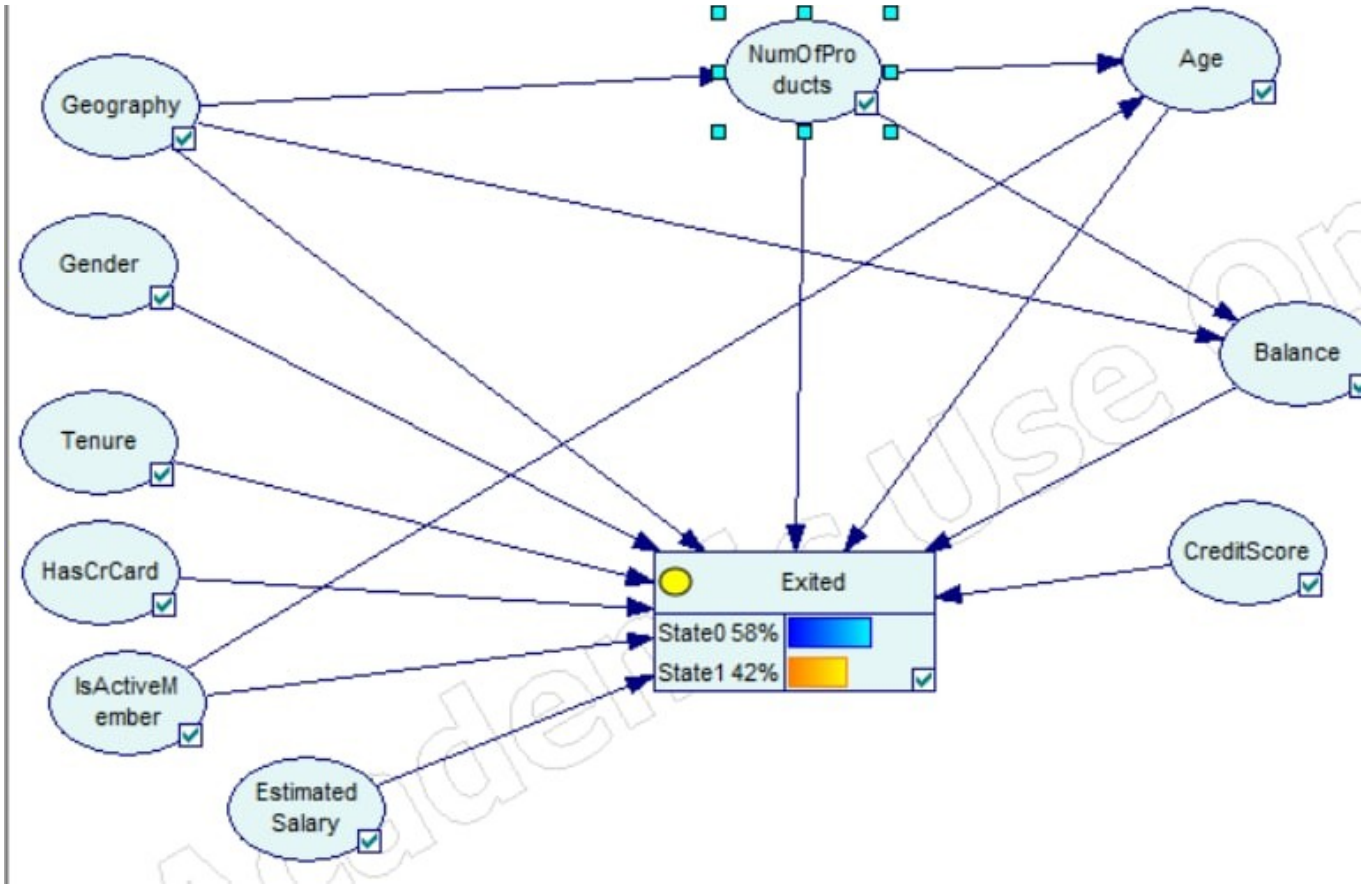
Результати тренування параметрів:



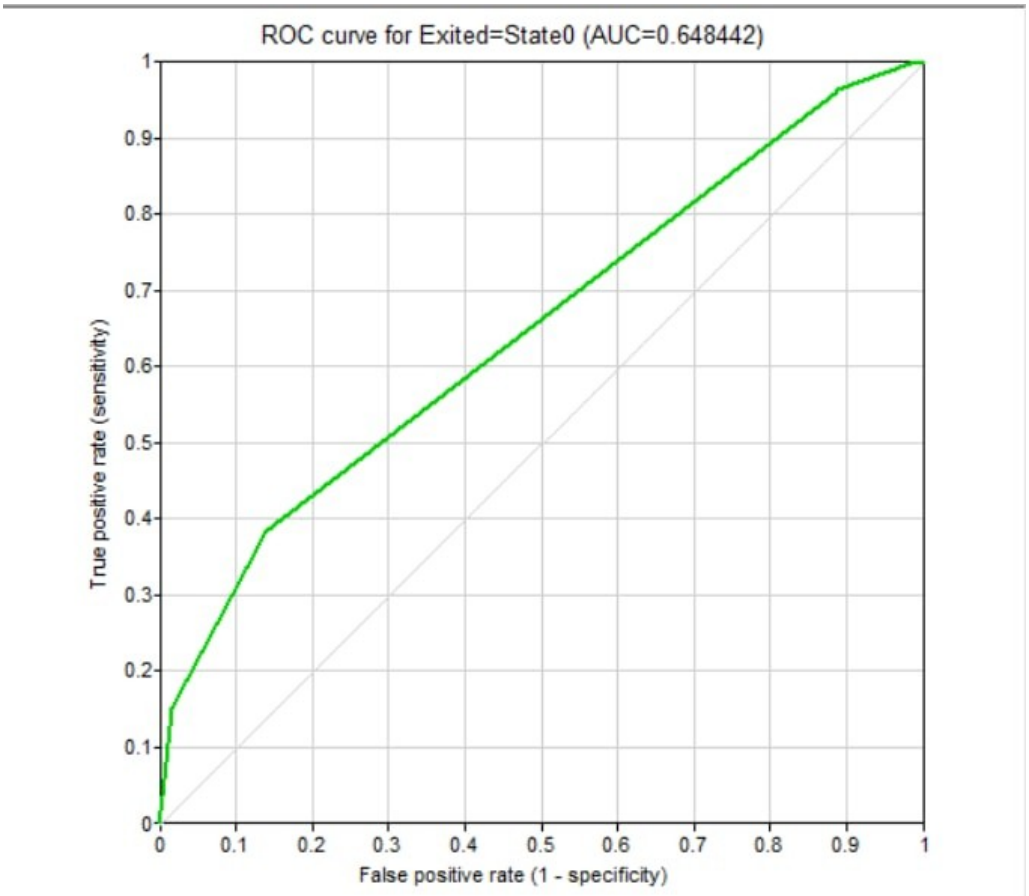
Алгоритм Greedy ThinkThinning має лише один параметр:
Максимальна кількість батьків (за замовчуванням 8) обмежує кількість батьків, яку може мати вузол. Оскільки розмір таблиць умовних ймовірностей вузла експоненціально зростає в порівнянні з кількістю батьків вузла, доцільно встановити обмеження на кількість батьків, щоб побудова мережі не вичерпувала всю доступну пам’ять комп’ютера.
Як ми бачимо, алгоритм має точність 81%, тобто правильно класифікує 812 випадків із 1002. Більша частина правильних випадків припадає на 0 класс, тобто на клієнтів, які не відмовились від послуг банку. Випадки ж 1 мають значно меншу точність 11%, тобто із 179 випадків лише 20 було класифіковано правильно.

Bayesian search
Алгоритм навчання структури байєсіанського пошуку є одним з найбільш ранніх і найпопулярніших алгоритмів, які використовуються. По суті, це слідує за процедурою підйому на гору (керується евристичною оцінкою, яка в GeNIe є функцією логарифмічної правдоподібності) із випадковими перезапусками.
Алгоритм створює ациклічний орієнтований граф, який отримує найвищий бал.
Оцінка пропорційна вірогідності даних з урахуванням структури, яка, якщо

припустити, що ми присвоюємо ту саму попередню ймовірність будь-якій структурі, пропорційна ймовірності структури з даними даними. Алгоритм створює на екрані текстове поле, яке містить налаштування всіх параметрів алгоритмів БС.



Результати валідації:
Exited = 0.810379 (812/1002)
State0 = 0.962333 (792/823)
State1 = 0.111732 (20/179)
ROC curve:

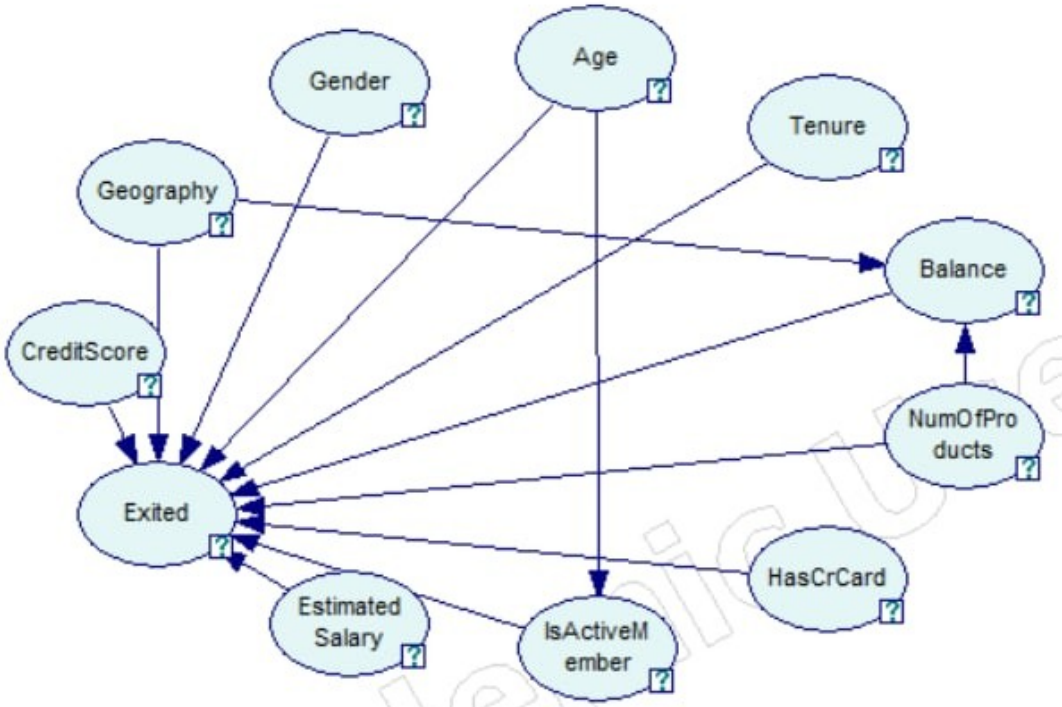


Як ми бачимо, алгоритм має точність 81%, тобто правильно класифікує 812 випадків із 1002. Більша частина правильних випадків припадає на 0 класс, тобто на клієнтів, які не відмовились від послуг банку. Випадки ж 1 мають значно меншу точність 11%, тобто із 179 випадків лише 20 було класифіковано правильно.

PC

Алгоритм навчання структури PC є одним з найперших і найпопулярніших алгоритмів, введених (Spirtes et al., 1993). Він використовує незалежності, які спостерігаються в даних (встановлені за допомогою класичних тестів незалежності), щоб зробити висновок про структуру, яка їх породила.

Алгоритм PC є єдиним алгоритмом вивчення структури в GeNIe, який дозволяє зберігати безперервні дані. Дані повинні відповідати розумному припущенню, що вони походять із багатовимірного нормального розподілу. Щоб перевірити це припущення, будь ласка, перевірте, чи гістограми кожної змінної близькі до нормального розподілу і що діаграми розсіювання кожної пари змінних показують приблизно лінійні співвідношення. Voortman & Druzdzel (2008) експериментально підтвердили, що алгоритм PC досить стійкий до припущення багатоваріантної нормальності.



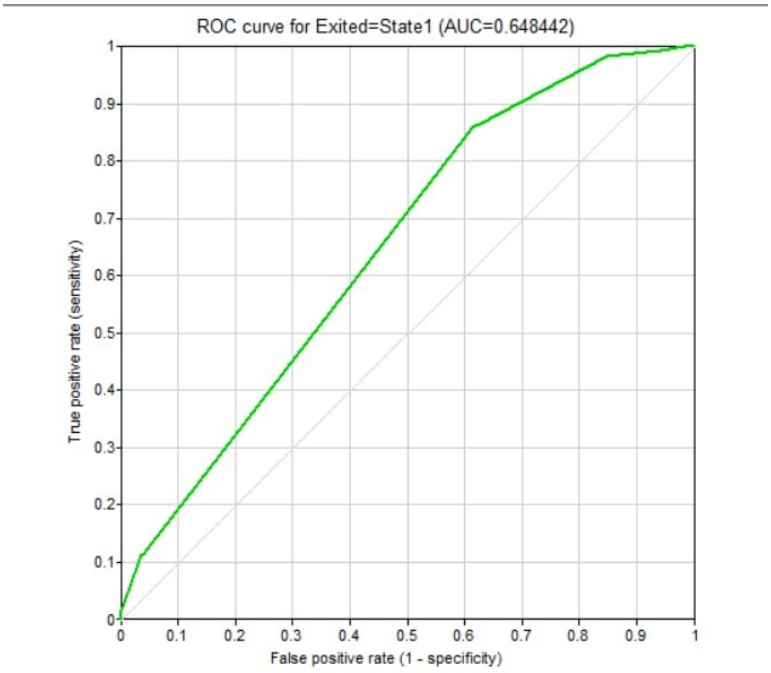
Input file: data_discret_learn.csv
Data rows: 8998
Elapsed time: 0.047s

Learning algorithm: PC
Algorithm parameters:
Max adjacency: 8
Significance: 0.05
Max search time: 0
Background knowledge was provided:
  forced arcs: 10

EM Log Likelihood: -95369.5

Результати валідації:
Exited = 0.810379 (812/1002)
  State0 = 0.962333 (792/823)
  State1 = 0.111732 (20/179)

ROC curve



Як ми бачимо, алгоритм має точність 81%, тобто правильно класифікує 812 випадків із 1002. Більша частина правильних випадків припадає на 0 класс, тобто на клієнтів, які не відмовились від послуг банку. Випадки ж 1 мають значно меншу точність 11%, тобто із 179 випадків лише 20 було класифіковано правильно.

Проаналізуємо отримані результати, для цього провалідуємо побудовані моделі через тестові данні та порівняємо точності моделей. Для Bayesian search

		Predicted	
		State0	State1
Act.	State0	**792**	31
	State1	159	**20**

Для Greedy ThinkThinning

		Predicted	
		State0	State1
Act.	State0	**792**	31
	State1	159	**20**

Для PC

		Predicted	
		State0	State1
Act.	State0	**792**	31
	State1	159	**20**

	Type 1 ERR	Type 2 ERR	OV.PERC.	AREA	GINI
Greedy ThinkThinning	31	159	0.81	0.65	0.3
PC	31	159	0.81	0.65	0.3
Bayesian search	31	159	0.81	0.65	0.3

Як ми бачимо, три результати для наших данних повністю співпали, три побудовані мережі прогнозують фінальний результат із однаковою кількістю похибок першого та другого роду. Це повністю співпадає із отриманниими результатами побудови, оскільки три мережі мали однакову точність 0.810379, а індекс GINI для трьох випадків становить приблизно 0.3

Практична частина - Дерева

1.CHAID
Автоматичне виявлення взаємодії хі-квадрат (CHAID) — це техніка дерева рішень, заснована на тестуванні скоригованої значущості (тестування Бонферроні). Методика була розроблена в Південній Африці і опублікована в 1980 році Гордоном В. Кассом, який захистив кандидатську дисертацію на цю тему. CHAID можна використовувати для прогнозування (аналогічно регресійному аналізу, ця

версія CHAID спочатку відома як XAID), а також для класифікації та для виявлення взаємодії між змінними. CHAID базується на офіційному розширенні процедур Сполучених Штатів Америки (Автоматичне виявлення взаємодії) і THAID (THeta Automatic Interaction Detection) 1960-х і 1970-х років, які, у свою чергу, були розширенням попередніх досліджень, у тому числі проведених у Великобританії в 1950-ті роки.

На практиці CHAID часто використовується в контексті прямого маркетингу для відбору груп споживачів і прогнозування того, як їхні реакції на одні змінні впливають на інші змінні, хоча інші ранні застосування були в області медичних і психіатричних досліджень.

Risk

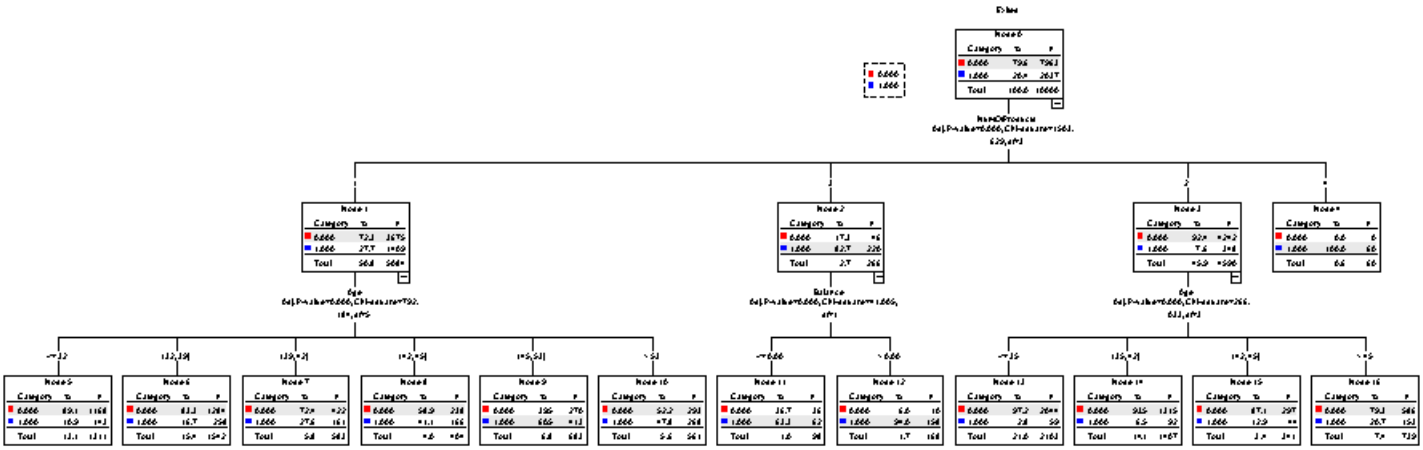
Method	Estimate	Std. Error
Resubstitution	.145	.004
Cross-Validation	.145	.004

Growing Method: CHAID
Dependent Variable: Exited

Classification

	Predicted		
Observed	0	1	Percent Correct
0	7747	216	97.3%
1	1229	808	39.7%
Overall Percentage	89.8%	10.2%	85.6%

Growing Method: CHAID
Dependent Variable: Exited



Для отриманих результатів збережемо побудовані вірогідності та розглянемо

Percantage оцінки із вказанним рівнем відсікання

Percantage:

95.00%	0.8109
90.00%	0.8109
85.00%	0.8326
80.00%	0.8326

Як ми бачимо, метод має точність 81-83%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 86%, проте таке рішення не є цілком виправданним, оскільки призводить до значного росту кількості помилок другого роду.

2. Exhaustive CHAID

Вичерпний CHAID є модифікацією CHAID, яка досліджує всі можливі розщеплення для кожного предиктора (Biggs et al., 1991). CRT — це сімейство методів, які максимізують однорідність всередині вузла (Breiman et al., 1984). Дерева QUEST обчислюються швидко, але метод доступний, лише якщо залежна змінна є номінальною.

Як ми бачимо, метод має точність 80-83%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 86%, проте таке рішення не є цілком виправданним, оскільки призводить до значного росту кількості помилок другого роду.

Risk

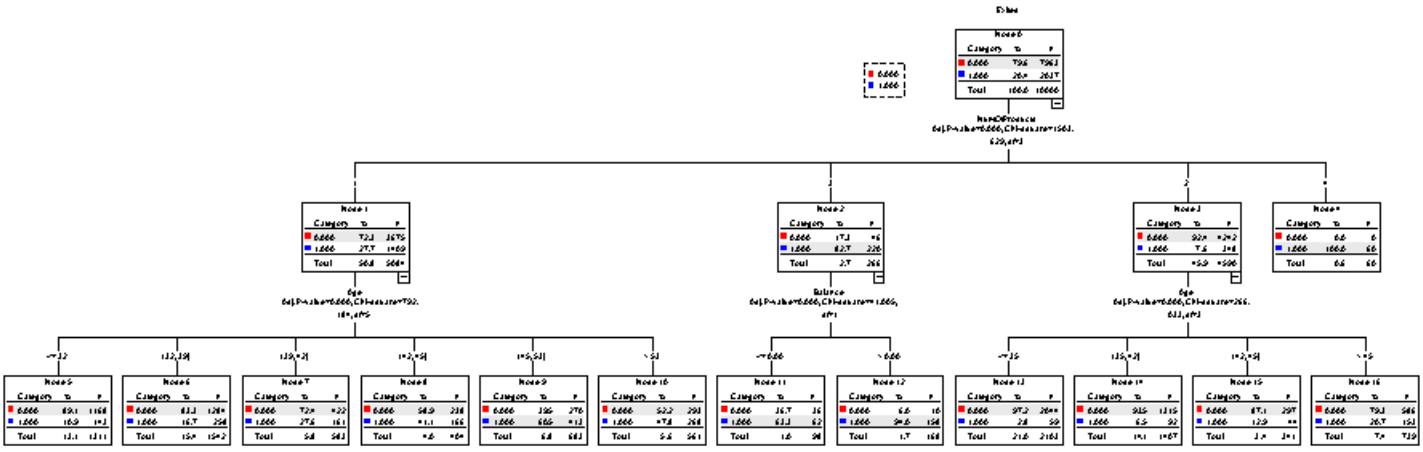
Method	Estimate	Std. Error
Resubstitution	.144	.004
Cross-Validation	.145	.004

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Exited

Classification

	Predicted		
Observed	0	1	Percent Correct
0	7747	216	97.3%
1	1229	808	39.7%
Overall Percentage	89.8%	10.2%	85.6%

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: Exited



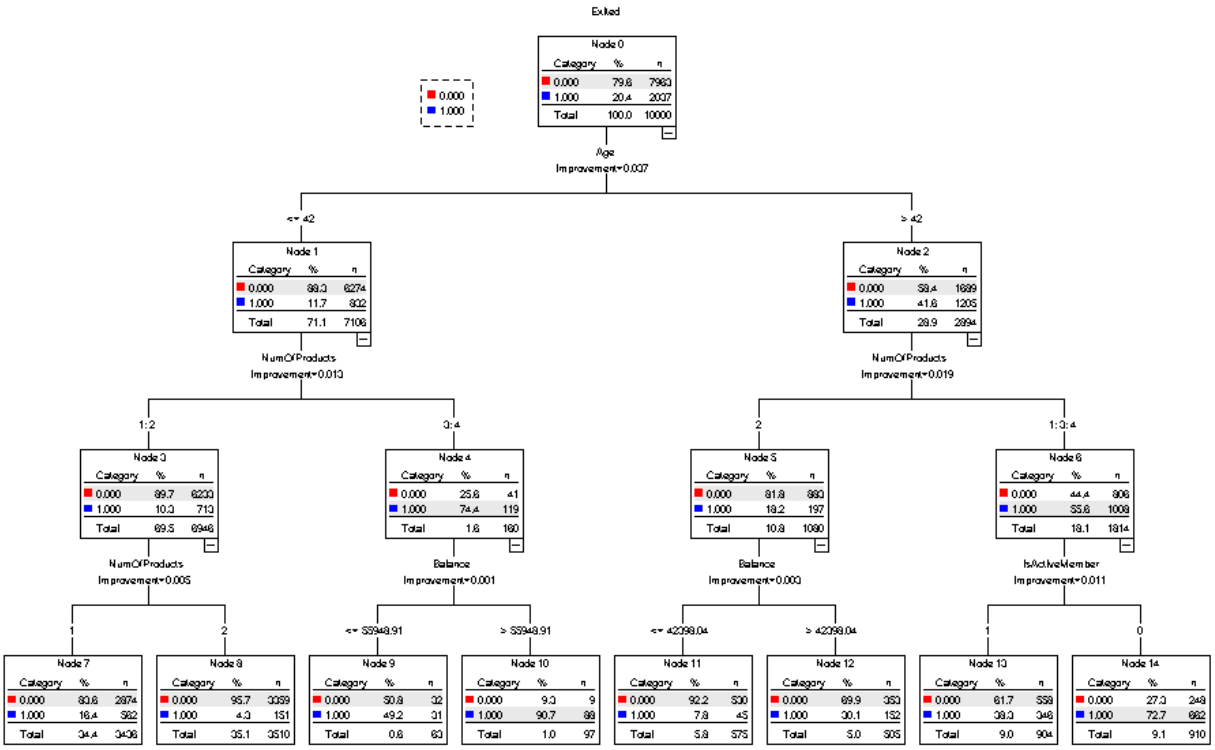
Percantage:

95.00%	0.8095
90.00%	0.8095
85.00%	0.8326
80.00%	0.8326

Як ми бачимо, метод має точність 80-83%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 86%, проте таке рішення не є цілком виправданним, оскільки призводить до значного рісту кількості помилок другого роду.

3. CRT

Дерева класифікації та регресії. CRT розбиває дані на сегменти, які є максимально однорідними щодо залежної змінної. Термінальний вузол, у якому всі випадки мають однакове значення для залежної змінної, є однорідним «чистим» вузлом.



Risk

Method	Estimate	Std. Error
Resubstitution	.147	.004
Cross-Validation	.150	.004

Growing Method: CRT
Dependent Variable: Exited

Classification

Observed	Predicted 0	Predicted 1	Percent Correct
0	7595	368	95.4%
1	1102	935	45.9%
Overall Percentage	87.0%	13.0%	85.3%

Growing Method: CRT
Dependent Variable: Exited

Percantage:

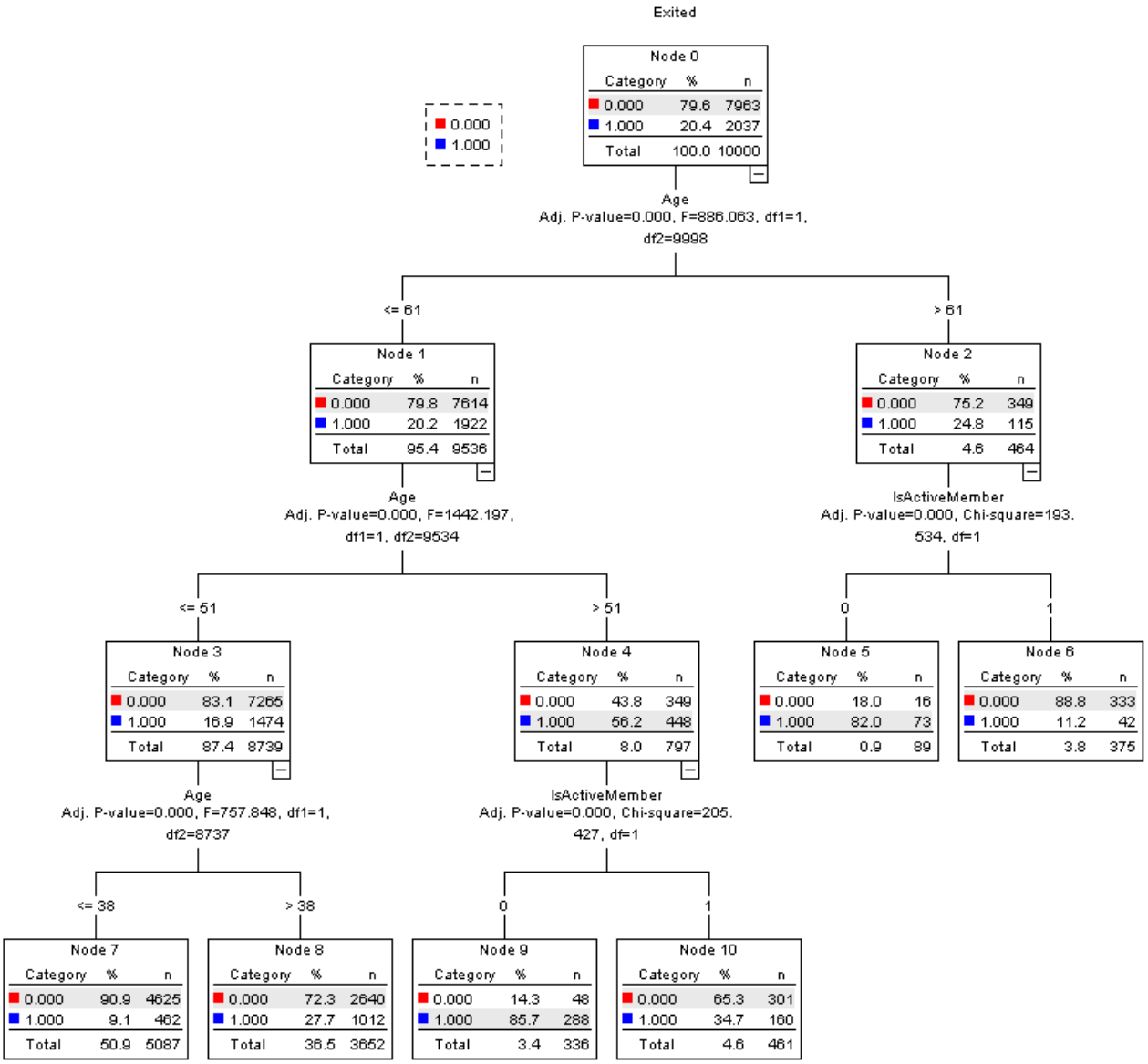
95.00%	0.7963
90.00%	0.8102
85.00%	0.8394
80.00%	0.8394

Як ми бачимо, метод має точність 79-83%. Зниження рівня відсікання може

підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 85%, проте таке рішення не є цілком виправданним, оскільки призводить до значного рісту кількості помилок другого роду.

4. QUEST

Швидке, неупереджене, ефективне статистичне дерево. Метод, який є швидким і уникає зміщення інших методів на користь предикторів з багатьма категоріями.



QUEST можна використати, тільки якщо залежна змінна є номінальною.

Percantage:

95.00%	0.7963
90.00%	0.8181
85.00%	0.8181
80.00%	0.8238

Як ми бачимо, метод має точність 79-82%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 85%, проте таке рішення не є цілком виправданним, оскільки призводить до значного рісту кількості помилок другого роду.

Risk

Method	Estimate	Std. Error
Resubstitution	.171	.004
Cross-Validation	.173	.004

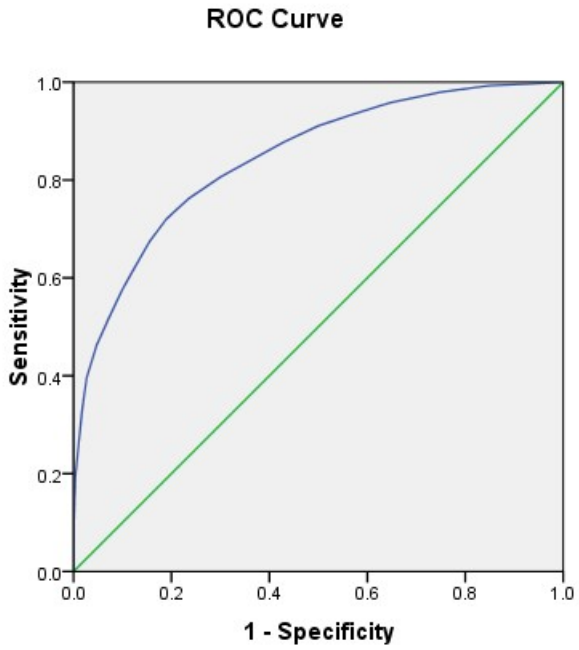
Growing Method: QUEST
Dependent Variable: Exited

Classification

Observed	Predicted		Percent Correct
	0	1	
0	7646	317	96.0%
1	1392	645	31.7%
Overall Percentage	90.4%	9.6%	82.9%

Growing Method: QUEST
Dependent Variable: Exited

Побудуємо відповідні ROC криві
1.CHAID



Diagonal segments are produced by ties.

Area Under the Curve

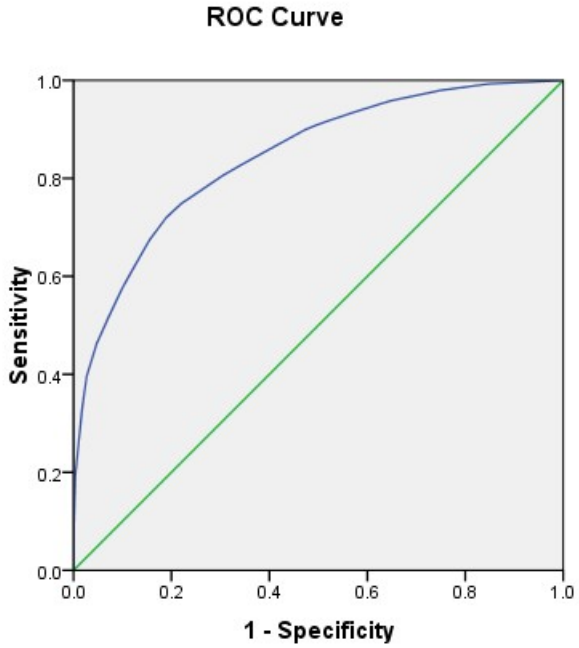
Test Result Variable(s):Predicted Probability for Exited=1

Area
.844

The test result variable(s): Predicted Probability for Exited=1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

Відповідне значення площі - 0.844264998075599

2. Exhaustive CHAID



Diagonal segments are produced by ties.

Area Under the Curve

Test Result Variable(s):Predicted Probability for Exited=1

Area
.844

The test result variable(s): Predicted Probability for Exited=1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

Відповідне значення площі — 0.8436873078488747

3.CRT

Area Under the Curve

Test Result Variable(s):Predicted Probability for Exited=1

Area
.845

The test result variable(s): Predicted Probability for Exited=1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.

Відповідне значення площі — 0.8449033826119341

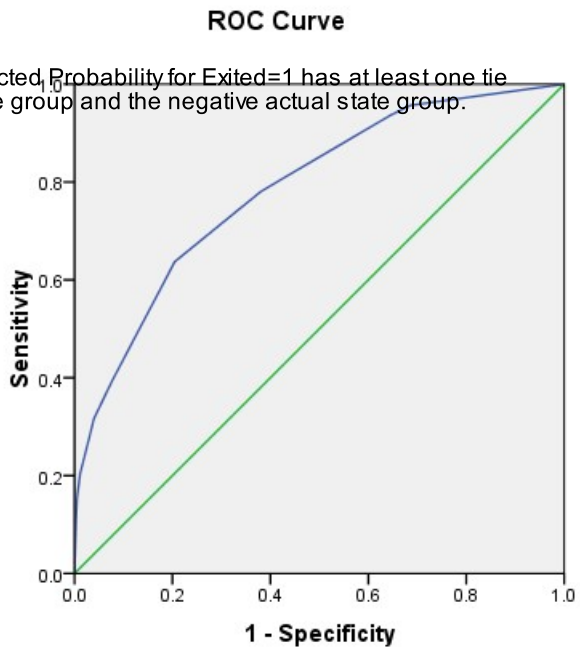
4.QUEST

Area Under the Curve

Test Result Variable(s):Predicted Probability for Exited=1

Area
.786

The test result variable(s): Predicted Probability for Exited=1 has at least one tie between the positive actual state group and the negative actual state group. Statistics may be biased.



Відповідне значення площі — 0.7857332738781864

Кількісну інтерпретацію ROC дає показник AUC (англ. area under ROC curve, площа під ROC-кривою) — площа, обмежена ROC-кривою і віссю частки помилкових позитивних класифікацій. Чим вище показник AUC, тим якісніше діє класифікатор, при цьому значення 0,5 демонструє непридатність обраного методу класифікації (відповідає звичайному вгадуванню).
У нашому випадку, показники AUC трохи менше 0.8, тобто модель показує себе досить непагано.

Таблиця отриманих результатів

	CHAID	Exhaustive CHAID	CRT	QUEST
95.00%				
Type I error	0	0	0	0
Type II error	1891	1905	2037	2037
Overrall Percantege	0.81	0.81	0.8	0.8
90.00%				
Type I error	30	30	56	27
Type II error	1891	1905	1884	1792
Overrall Percantege	0.81	0.81	0.81	0.82
85.00%				
Type I error	30	30	56	27
Type II error	1644	1644	1550	1792
Overrall Percantege	0.83	0.83	0.84	0.82
80.00%				
Type I error	30	30	56	43
Type II error	1644	1644	1550	1719
Overrall Percantege	0.83	0.83	0.84	0.82

	CHAID	EX. CHAID	CRT	QUEST
ROC Area	0.84	0.84	0.84	0.79
GINI INDEX	0.69	0.69	0.69	0.57

Як ми бачимо, найкращу точність демонструє алгоритм побудови CRT, який при рівнях відсікання 80-85% показує точність у 84% при найменшій кількості помилок першого та другого роду.

Практична частина - Регресія

1. Логістична регресія методом Enter
Процедура вибору змінної, в якій всі змінні в блоці вводяться за один крок.
Поетапно . На кожному кроці вводиться незалежна змінна, яка не входить до рівняння, яка має найменшу ймовірність F, якщо ця ймовірність достатньо мала.

Classification Table[a]

Observed			Predicted		
			Exited		Percentage Correct
			0	1	
Step 1	Exited	0	7721	242	97.0
		1	1682	355	17.4
	Overall Percentage				80.8

a. The cut value is .500

Побудована математична модель:

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CreditScore	-.004	.000	358.467	1	.000	.996
	Geography	-.015	.032	.214	1	.643	.985
	Gender	-.638	.053	146.109	1	.000	.528
	Age	.058	.002	631.207	1	.000	1.060
	Tenure	-.039	.009	19.019	1	.000	.962
	Balance	.000	.000	60.998	1	.000	1.000
	NumOfProducts	-.268	.045	36.038	1	.000	.765
	HasCrCard	-.182	.056	10.479	1	.001	.834
	IsActiveMember	-1.078	.056	371.220	1	.000	.340
	EstimatedSalary	.000	.000	3.638	1	.056	1.000

a. Variable(s) entered on step 1: CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary.

Percantage:

95.00%	0.7963
90.00%	0.7964
85.00%	0.7965
80.00%	0.797

Як ми бачимо, метод має точність 79%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 82%, проте таке рішення не є цілком виправданним, оскільки призводить до значного росту кількості помилок другого роду.

2. Логістична регресія методом Forward Stepwise Conditional

Метод поетапного відбору з вхідним тестуванням на основі значущості статистики оцінок і тестуванням на видалення на основі статистики ймовірності-відношення на основі оцінок умовних параметрів.

Classification Table[a]

Observed			Predicted		
			Exited		Percentage Correct
			0	1	
Step 1	Exited	0	7963	0	100.0
		1	2037	0	.0
	Overall Percentage				79.6
Step 2	Exited	0	7789	174	97.8
		1	1942	95	4.7
	Overall Percentage				78.8
Step 3	Exited	0	7805	158	98.0
		1	1760	277	13.6
	Overall Percentage				80.8
Step 4	Exited	0	7753	210	97.4
		1	1731	306	15.0
	Overall Percentage				80.6
Step 5	Exited	0	7709	254	96.8
		1	1688	349	17.1
	Overall Percentage				80.6
Step 6	Exited	0	7718	245	96.9
		1	1681	356	17.5
	Overall Percentage				80.7
Step 7	Exited	0	7725	238	97.0
		1	1683	354	17.4
	Overall Percentage				80.8
Step 8	Exited	0	7731	232	97.1
		1	1681	356	17.5
	Overall Percentage				80.9

a. The cut value is .500

Variables not in the Equation[a]

			Score	df	Sig.
Step 1	Variables	Geography	1.110	1	.292
		Gender	138.694	1	.000
		Age	435.052	1	.000
		Tenure	11.352	1	.001
		Balance	103.908	1	.000
		NumOfProducts	53.564	1	.000
		HasCrCard	6.513	1	.011
		IsActiveMember	269.917	1	.000
		EstimatedSalary	.785	1	.376
Step 2	Variables	Geography	2.117	1	.146
		Gender	164.216	1	.000
		Tenure	26.646	1	.000
		Balance	75.880	1	.000
		NumOfProducts	90.541	1	.000
		HasCrCard	17.292	1	.000
		IsActiveMember	402.927	1	.000
		EstimatedSalary	7.314	1	.007
Step 3	Variables	Geography	1.430	1	.232
		Gender	153.471	1	.000
		Tenure	31.657	1	.000
		Balance	72.739	1	.000
		NumOfProducts	81.532	1	.000
		HasCrCard	19.146	1	.000
		EstimatedSalary	8.080	1	.004
Step 4	Variables	Geography	.470	1	.493
		Tenure	25.140	1	.000
		Balance	82.406	1	.000
		NumOfProducts	76.571	1	.000
		HasCrCard	15.731	1	.000
		EstimatedSalary	6.384	1	.012
Step 5	Variables	Geography	2.760	1	.097
		Tenure	28.246	1	.000
		NumOfProducts	49.142	1	.000
		HasCrCard	17.079	1	.000
		EstimatedSalary	8.962	1	.003
Step 6	Variables	Geography	.710	1	.400
		Tenure	21.964	1	.000
		HasCrCard	12.920	1	.000
		EstimatedSalary	5.205	1	.023
Step 7	Variables	Geography	.436	1	.509
		HasCrCard	11.050	1	.001
		EstimatedSalary	4.161	1	.041
Step 8	Variables	Geography	.298	1	.585
		EstimatedSalary	3.723	1	.054

a. Residual Chi-Squares are not computed because of redundancies.

Percantage:

95.00%	0.7964
90.00%	0.7964
85.00%	0.7966
80.00%	0.7972

Як ми бачимо, метод має точність 79%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 82%, проте таке рішення не є цілком виправданним, оскільки

призводить до значного рісту кількості помилок другого роду.

3. Логістична регресія методом Backward Stepwise Conditional

Classification Table[a]

Observed			Predicted		
			Exited		Percentage Correct
			0	1	
Step 1	Exited	0	7721	242	97.0
		1	1682	355	17.4
	Overall Percentage				80.8
Step 2	Exited	0	7727	236	97.0
		1	1684	353	17.3
	Overall Percentage				80.8

a. The cut value is .500

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	8921.604[a]	.390	.520
2	8921.819[a]	.390	.520

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CreditScore	-.004	.000	358.467	1	.000	.996
	Geography	-.015	.032	.214	1	.643	.985
	Gender	-.638	.053	146.109	1	.000	.528
	Age	.058	.002	631.207	1	.000	1.060
	Tenure	-.039	.009	19.019	1	.000	.962
	Balance	.000	.000	60.998	1	.000	1.000
	NumOfProducts	-.268	.045	36.038	1	.000	.765
	HasCrCard	-.182	.056	10.479	1	.001	.834
	IsActiveMember	-1.078	.056	371.220	1	.000	.340
	EstimatedSalary	.000	.000	3.638	1	.056	1.000
Step 2[a]	CreditScore	-.004	.000	382.711	1	.000	.996
	Gender	-.639	.053	146.589	1	.000	.528
	Age	.058	.002	635.756	1	.000	1.060
	Tenure	-.039	.009	19.181	1	.000	.961
	Balance	.000	.000	61.023	1	.000	1.000
	NumOfProducts	-.270	.044	37.029	1	.000	.763
	HasCrCard	-.183	.056	10.592	1	.001	.833
	IsActiveMember	-1.078	.056	371.493	1	.000	.340
	EstimatedSalary	.000	.000	3.721	1	.054	1.000

a. Variable(s) entered on step 1: CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary.

Percantage:

95.00%	0.7964
90.00%	0.7964
85.00%	0.7965
80.00%	0.7971

Як ми бачимо, метод має точність 79%. Зниження рівня відсікання може підвищити

точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 82%, проте таке рішення не є цілком виправданним, оскільки призводить до значного рісту кількості помилок другого роду.

4. Логістична регресія методом Forward LR

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	10132.378[a]	.311	.415
2	9656.777[b]	.343	.458
3	9245.018[c]	.370	.493
4	9091.644[c]	.379	.506
5	9008.523[c]	.385	.513
6	8958.477[c]	.388	.517
7	8936.481[c]	.389	.519
8	8925.543[c]	.390	.520

a. Estimation terminated at iteration number 2 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

c. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step…	CreditScore	-.002	.000	2.931E3	1	.000	.998
Step 2[b]	CreditScore	-.005	.000	1.201E3	1	.000	.995
	Age	.044	.002	456.166	1	.000	1.045
Step 3[c]	CreditScore	-.005	.000	1.064E3	1	.000	.995
	Age	.054	.002	587.678	1	.000	1.055
	IsActiveMember	-1.073	.055	384.202	1	.000	.342
Step 4[d]	CreditScore	-.004	.000	874.233	1	.000	.996
	Gender	-.639	.052	151.007	1	.000	.528
	Age	.056	.002	616.572	1	.000	1.057
	IsActiveMember	-1.070	.055	374.414	1	.000	.343
Step 5[e]	CreditScore	-.005	.000	926.186	1	.000	.995
	Gender	-.663	.052	160.423	1	.000	.515
	Age	.054	.002	572.982	1	.000	1.055
	Balance	.000	.000	81.621	1	.000	1.000
	IsActiveMember	-1.068	.055	371.718	1	.000	.344
Step 6[f]	CreditScore	-.004	.000	532.425	1	.000	.996
	Gender	-.654	.053	154.837	1	.000	.520
	Age	.056	.002	609.640	1	.000	1.058
	Balance	.000	.000	54.301	1	.000	1.000
	NumOfProducts	-.307	.044	48.735	1	.000	.736
	IsActiveMember	-1.065	.056	366.067	1	.000	.345
Step 7[g]	CreditScore	-.004	.000	458.344	1	.000	.996
	Gender	-.644	.053	149.590	1	.000	.525
	Age	.057	.002	624.158	1	.000	1.059
	Tenure	-.042	.009	21.908	1	.000	.959
	Balance	.000	.000	57.842	1	.000	1.000
	NumOfProducts	-.288	.044	42.704	1	.000	.749
	IsActiveMember	-1.074	.056	369.918	1	.000	.342
Step 8[h]	CreditScore	-.004	.000	415.935	1	.000	.996
	Gender	-.640	.053	147.383	1	.000	.527
	Age	.058	.002	632.234	1	.000	1.060
	Tenure	-.040	.009	20.081	1	.000	.961
	Balance	.000	.000	59.278	1	.000	1.000
	NumOfProducts	-.278	.044	39.568	1	.000	.757
	HasCrCard	-.186	.056	11.034	1	.001	.830
	IsActiveMember	-1.077	.056	371.101	1	.000	.341

a. Variable(s) entered on step 1: CreditScore.

b. Variable(s) entered on step 2: Age.

c. Variable(s) entered on step 3: IsActiveMember.

d. Variable(s) entered on step 4: Gender.

e. Variable(s) entered on step 5: Balance.

f. Variable(s) entered on step 6: NumOfProducts.

g. Variable(s) entered on step 7: Tenure.

h. Variable(s) entered on step 8: HasCrCard.

Classification Table[a]

			Predicted		
			Exited		
	Observed		0	1	Percentage Correct
Step 1	Exited	0	7963	0	100.0
		1	2037	0	.0
	Overall Percentage				79.6
Step 2	Exited	0	7789	174	97.8
		1	1942	95	4.7
	Overall Percentage				78.8
Step 3	Exited	0	7805	158	98.0
		1	1760	277	13.6
	Overall Percentage				80.8
Step 4	Exited	0	7753	210	97.4
		1	1731	306	15.0
	Overall Percentage				80.6
Step 5	Exited	0	7709	254	96.8
		1	1688	349	17.1
	Overall Percentage				80.6
Step 6	Exited	0	7718	245	96.9
		1	1681	356	17.5
	Overall Percentage				80.7
Step 7	Exited	0	7725	238	97.0
		1	1683	354	17.4
	Overall Percentage				80.8
Step 8	Exited	0	7731	232	97.1
		1	1681	356	17.5
	Overall Percentage				80.9

a. The cut value is .500

Percantage:

95.00%	0.7964
90.00%	0.7964
85.00%	0.7966
80.00%	0.7972

Як ми бачимо, метод має точність 79%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 82%, проте таке рішення не є цілком виправданним, оскільки призводить до значного росту кількості помилок другого роду.

5. Логістична регресія методом Backward LR

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	8921.604[a]	.390	.520
2	8921.819[a]	.390	.520

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Exited		
	Observed		0	1	Percentage Correct
Step 1	Exited	0	7721	242	97.0
		1	1682	355	17.4
	Overall Percentage				80.8
Step 2	Exited	0	7727	236	97.0
		1	1684	353	17.3
	Overall Percentage				80.8

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CreditScore	-.004	.000	358.467	1	.000	.996
	Geography	-.015	.032	.214	1	.643	.985
	Gender	-.638	.053	146.109	1	.000	.528
	Age	.058	.002	631.207	1	.000	1.060
	Tenure	-.039	.009	19.019	1	.000	.962
	Balance	.000	.000	60.998	1	.000	1.000
	NumOfProducts	-.268	.045	36.038	1	.000	.765
	HasCrCard	-.182	.056	10.479	1	.001	.834
	IsActiveMember	-1.078	.056	371.220	1	.000	.340
	EstimatedSalary	.000	.000	3.638	1	.056	1.000
Step 2[a]	CreditScore	-.004	.000	382.711	1	.000	.996
	Gender	-.639	.053	146.589	1	.000	.528
	Age	.058	.002	635.756	1	.000	1.060
	Tenure	-.039	.009	19.181	1	.000	.961
	Balance	.000	.000	61.023	1	.000	1.000
	NumOfProducts	-.270	.044	37.029	1	.000	.763
	HasCrCard	-.183	.056	10.592	1	.001	.833
	IsActiveMember	-1.078	.056	371.493	1	.000	.340
	EstimatedSalary	.000	.000	3.721	1	.054	1.000

a. Variable(s) entered on step 1: CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary.

Percantage:

95.00%	0.7964
90.00%	0.7964
85.00%	0.7965
80.00%	0.7971

Як ми бачимо, метод має точність 79%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 82%, проте таке рішення не є цілком виправданним, оскільки призводить до значного рісту кількості помилок другого роду.

6. Логістична регресія методом Forward Wald

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	10132.378[a]	.311	.415
2	9656.777[b]	.343	.458
3	9245.018[c]	.370	.493
4	9091.644[c]	.379	.506
5	9008.523[c]	.385	.513
6	8958.477[c]	.388	.517
7	8936.481[c]	.389	.519
8	8925.543[c]	.390	.520

a. Estimation terminated at iteration number 2 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

c. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

Observed			Predicted		
			Exited		Percentage Correct
			0	1	
Step 1	Exited	0	7963	0	100.0
		1	2037	0	.0
	Overall Percentage				79.6
Step 2	Exited	0	7789	174	97.8
		1	1942	95	4.7
	Overall Percentage				78.8
Step 3	Exited	0	7805	158	98.0
		1	1760	277	13.6
	Overall Percentage				80.8
Step 4	Exited	0	7753	210	97.4
		1	1731	306	15.0
	Overall Percentage				80.6
Step 5	Exited	0	7709	254	96.8
		1	1688	349	17.1
	Overall Percentage				80.6
Step 6	Exited	0	7718	245	96.9
		1	1681	356	17.5
	Overall Percentage				80.7
Step 7	Exited	0	7725	238	97.0
		1	1683	354	17.4
	Overall Percentage				80.8
Step 8	Exited	0	7731	232	97.1
		1	1681	356	17.5
	Overall Percentage				80.9

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step...[a]	CreditScore	-.002	.000	2.931E3	1	.000	.998
Step 2[b]	CreditScore	-.005	.000	1.201E3	1	.000	.995
	Age	.044	.002	456.166	1	.000	1.045
Step 3[c]	CreditScore	-.005	.000	1.064E3	1	.000	.995
	Age	.054	.002	587.678	1	.000	1.055
	IsActiveMember	-1.073	.055	384.202	1	.000	.342
Step 4[d]	CreditScore	-.004	.000	874.233	1	.000	.996
	Gender	-.639	.052	151.007	1	.000	.528
	Age	.056	.002	616.572	1	.000	1.057
	IsActiveMember	-1.070	.055	374.414	1	.000	.343
Step 5[e]	CreditScore	-.005	.000	926.186	1	.000	.995
	Gender	-.663	.052	160.423	1	.000	.515
	Age	.054	.002	572.982	1	.000	1.055
	Balance	.000	.000	81.621	1	.000	1.000
	IsActiveMember	-1.068	.055	371.718	1	.000	.344
Step 6[f]	CreditScore	-.004	.000	532.425	1	.000	.996
	Gender	-.654	.053	154.837	1	.000	.520
	Age	.056	.002	609.640	1	.000	1.058
	Balance	.000	.000	54.301	1	.000	1.000
	NumOfProducts	-.307	.044	48.735	1	.000	.736
	IsActiveMember	-1.065	.056	366.067	1	.000	.345
Step 7[g]	CreditScore	-.004	.000	458.344	1	.000	.996
	Gender	-.644	.053	149.590	1	.000	.525
	Age	.057	.002	624.158	1	.000	1.059
	Tenure	-.042	.009	21.908	1	.000	.959
	Balance	.000	.000	57.842	1	.000	1.000
	NumOfProducts	-.288	.044	42.704	1	.000	.749
	IsActiveMember	-1.074	.056	369.918	1	.000	.342
Step 8[h]	CreditScore	-.004	.000	415.935	1	.000	.996
	Gender	-.640	.053	147.383	1	.000	.527
	Age	.058	.002	632.234	1	.000	1.060
	Tenure	-.040	.009	20.081	1	.000	.961
	Balance	.000	.000	59.278	1	.000	1.000
	NumOfProducts	-.278	.044	39.568	1	.000	.757
	HasCrCard	-.186	.056	11.034	1	.001	.830
	IsActiveMember	-1.077	.056	371.101	1	.000	.341

Percantage:

95.00%	0.7964
90.00%	0.7964
85.00%	0.7966
80.00%	0.7972

Як ми бачимо, метод має точність 79%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 82%, проте таке рішення не є цілком виправданним, оскільки призводить до значного росту кількості помилок другого роду.

7.Логістична регресія методом Backward Wald

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	CreditScore	-.004	.000	358.467	1	.000	.996
	Geography	-.015	.032	.214	1	.643	.985
	Gender	-.638	.053	146.109	1	.000	.528
	Age	.058	.002	631.207	1	.000	1.060
	Tenure	-.039	.009	19.019	1	.000	.962
	Balance	.000	.000	60.998	1	.000	1.000
	NumOfProducts	-.268	.045	36.038	1	.000	.765
	HasCrCard	-.182	.056	10.479	1	.001	.834
	IsActiveMember	-1.078	.056	371.220	1	.000	.340
	EstimatedSalary	.000	.000	3.638	1	.056	1.000
Step 2[a]	CreditScore	-.004	.000	382.711	1	.000	.996
	Gender	-.639	.053	146.589	1	.000	.528
	Age	.058	.002	635.756	1	.000	1.060
	Tenure	-.039	.009	19.181	1	.000	.961
	Balance	.000	.000	61.023	1	.000	1.000
	NumOfProducts	-.270	.044	37.029	1	.000	.763
	HasCrCard	-.183	.056	10.592	1	.001	.833
	IsActiveMember	-1.078	.056	371.493	1	.000	.340
	EstimatedSalary	.000	.000	3.721	1	.054	1.000

a. Variable(s) entered on step 1: CreditScore, Geography, Gender, Age, Tenure, Balance, NumOfProducts, HasCrCard, IsActiveMember, EstimatedSalary.

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	8921.604[a]	.390	.520
2	8921.819[a]	.390	.520

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			Exited		
	Observed		0	1	Percentage Correct
Step 1	Exited	0	7721	242	97.0
		1	1682	355	17.4
	Overall Percentage				80.8
Step 2	Exited	0	7727	236	97.0
		1	1684	353	17.3
	Overall Percentage				80.8

a. The cut value is .500

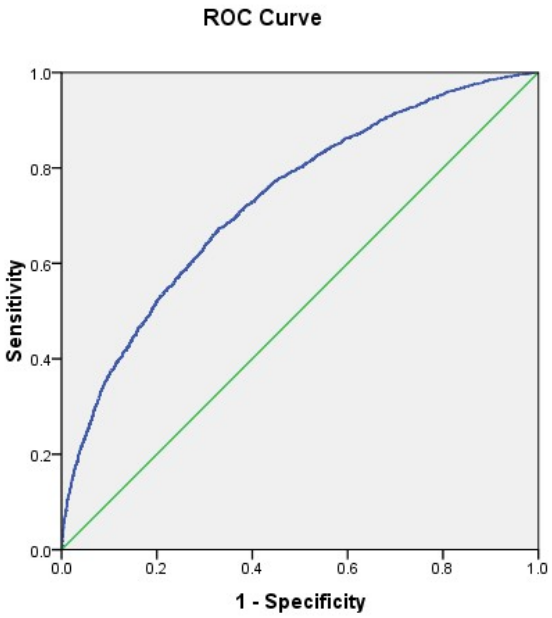
Percantage:

95.00%	0.7964
90.00%	0.7964
85.00%	0.7965
80.00%	0.7971

Як ми бачимо, метод має точність 79%. Зниження рівня відсікання може підвищити точність, наприклад при рівні у 50% процент правильно классифікованих випадків становить приблизно 82%, проте таке рішення не є цілком виправданним, оскільки призводить до значного росту кількості помилок другого роду.

ROC Curves:

1.Enter



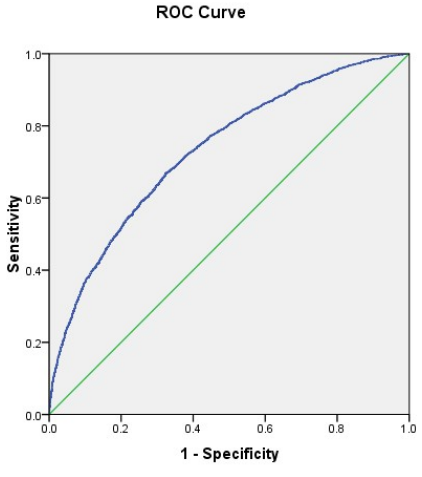
0.7320209676183375

Area Under the Curve

Test Result Variable(s):Predicted probability

Area
.732

2.Forward Ward



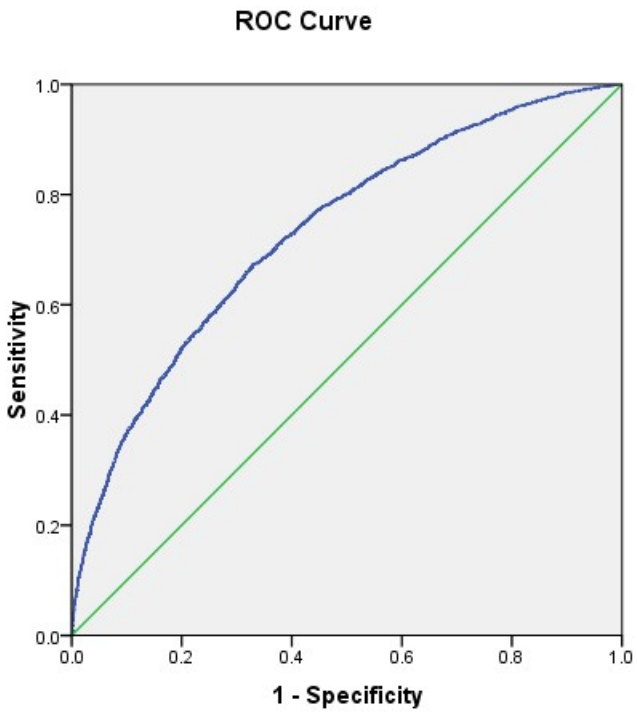
Area Under the Curve

Test Result Variable(s):Predicted probability

Area
.733

0.7325451765717376

3.Forward Conditional



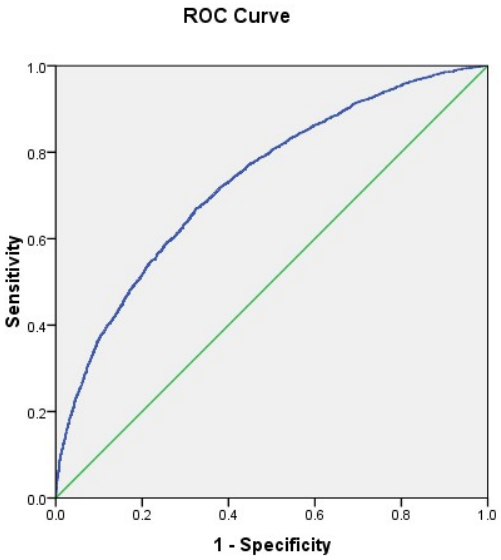
Area Under the Curve

Test Result Variable(s):Predicted probability

Area
.732

0.7320209676183375

4.Forward LR



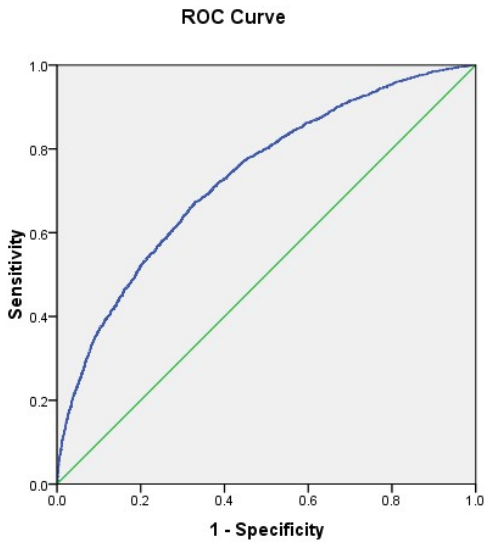
Area Under the Curve

Test Result Variable(s):Predicted probability

Area
.733

0.7325451765717376

5.Backward Ward



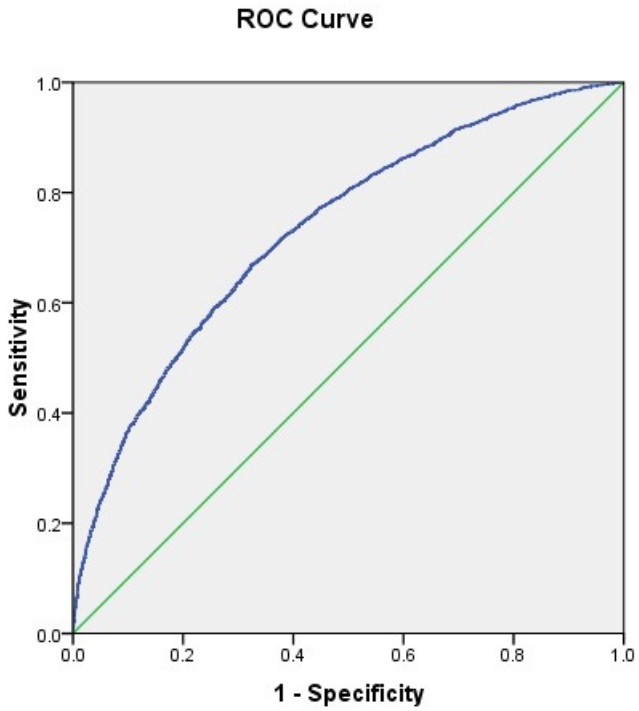
Area Under the Curve

Test Result Variable(s):Predicted probability

Area
.732

0.7320209676183375

6.Backward Conditional



Area Under the Curve

Test Result Variable(s):Predicted probability

Area
.733

0.7325451765717376

	Enter	Forward Ward	Forward Conditional	Forward LR	Backward Ward	Backward Conditional	Backward LR
95.00%							
Type I error	0	0	0	0	0	0	0
Type II error	2037	2037	2037	2037	2037	2037	2037
Overrall Percantege	0.8	0.8	0.8	0.8	0.8	0.8	0.8
90.00%							
Type I error	2	2	2	2	2	2	2
Type II error	2036	2036	2036	2036	2036	2036	2036
Overrall Percantege	0.8	0.8	0.8	0.8	0.8	0.8	0.8
85.00%							
Type I error	2	2	2	2	2	2	2
Type II error	2033	2033	2033	2033	2033	2033	2033
Overrall Percantege	0.8	0.8	0.8	0.8	0.8	0.8	0.8
80.00%							
Type I error	9	9	9	9	9	9	9
Type II error	2021	2022	2022	2022	2022	2022	2022
Overrall Percantege	0.8	0.8	0.8	0.8	0.8	0.8	0.8

	Enter	Forward Ward	Forward Conditional	Forward LR	Backward Ward	Backward Conditional	Backward LR
ROC Area	0.73	0.73	0.73	0.73	0.73	0.73	0.73
GINI INDEX	0.46	0.47	0.46	0.47	0.46	0.47	0.46

Як ми бачимо, усі результати для наших данних повністю співпали, використані алгоритми прогнозують фінальний результат із однаковою кількістю похибок першого та другого роду. Це повністю співпадає із отриманниими результатами побудови, оскільки три мережі мали однакову точність 0.8, а індекс GINI для усіх випадків становить приблизно 0.73

Таблиця отриманих результатів для дерев та регресіі

	95.00%	Type I error	Type II error	Overrall Percantege	90.00 %	Type I error	Type II error
CHAID		0	1891	0.81		30	1891
Exhaustive CHAID		0	1905	0.81		30	1905
CRT		0	2037	0.8		56	1884
QUEST		0	2037	0.8		27	1792
Enter		0	2037	0.8		2	2036
Forward Ward		0	2037	0.8		2	2036
Forward Conditional		0	2037	0.8		2	2036
Forward LR		0	2037	0.8		2	2036
Backward Ward		0	2037	0.8		2	2036
Backward Conditional		0	2037	0.8		2	2036
Backward LR		0	2037	0.8		2	2036

Overrall Percantege	85.00%	Type I error	Type II error	Overrall Percantege	80.00%	Type I error	Type II error	Overrall Percantege
0.81		30	1644	0.83		30	1644	0.83
0.81		30	1644	0.83		30	1644	0.83
0.81		56	1550	0.84		56	1550	0.84
0.82		27	1792	0.82		43	1719	0.82
0.8		2	2033	0.8		9	2021	0.8
0.8		2	2033	0.8		9	2022	0.8
0.8		2	2033	0.8		9	2022	0.8
0.8		2	2033	0.8		9	2022	0.8
0.8		2	2033	0.8		9	2022	0.8
0.8		2	2033	0.8		9	2022	0.8
0.8		2	2033	0.8		9	2022	0.8

	CHAID	EX. CHAID	CRT	QUEST	Enter
ROC Area	0.84	0.84	0.84	0.79	0.73
GINI INDEX	0.69	0.69	0.69	0.57	0.46

Forward Ward	Forward Conditional	Forward LR	Backward Ward	Backward Conditional	Backward LR
0.73	0.73	0.73	0.73	0.73	0.73
0.46	0.46	0.47	0.46	0.47	0.46

Серед усіх моделей найкращу точність має модель побудована методом CRT

– 0.84, що є дійсно дуже хорошим результатом. Методи CHAID та Exhaustive CHAID дали також хорошу однакову точність – 0.83. Найгіршу, але також хорошу, точність серед розглянутих методів дав метод QUEST.
Якщо порівнювати значення індексу GINI, то маємо, що найбільше значення має модель побудована за алгоритмоми CRT, CHAID та Exhaustive CHAID – 0.69, а найменше значення має модель побудована за регресійними алгоритмоми – 0.73.

Висновки

При дослідженні данного датасету ми використовували різні методи прогнозування, обробки та аналізу данних, а саме: Мережі Байєса, Дерева Рішеннь та Регрессійні Моделі. Ми побудували моделі використовуючи різні методи та підходи, намагаючись навчити моделі якнайточніше вирішувати завдання классифікації на основі певного датасету. У цілому різні моделі повернули приблизно однакову точність, тому очевидним є можливості комбінування їх для рішення real-case завдань із врахуванням особливостей кожних завданнь. Проте для кожного із єтапів ми розглянули фаворитів, а саме для мереж був обраний алгоритм Greedy ThinkThinning, для дерев CRT, для регресіі Enter

Огляд найкращих результатів на кожному із єтапів

	Type 1 ERR	Type 2 ERR	OV.PERC.	AREA	GINI
Greedy ThinkThinning	31	159	0.81	0.65	0.3
CRT – 85%	56	1550	0.84	0.84	0.69
Enter – 85%	9	2021	0.8	0.73	0.46

У нашому випадку, найкращу точність показали дерева рішеннь, які у середньому дали 83% точності прогнозу. Максимальний результат — метод CRT із 80%-85% порогом відсікання. За таких умов його точність була 0.8394.
Проте, аналізуючи отримані результати, очевидним є факт превалювання помилок другого роду та їх великої кількості. Це означає, що маючи набір данних, який характеризує кліента, який відмовився від послуг банку, модель не здатна адекватно классифікувати його як Відмовника. Така ситуація актуальна для усіх побудованих моделей. Така особливість системи ставить під сумнів її працездатність, оскільки точність для приблизно рівномірного датасету буде скоріш за все посередньою. Виправити ситуацію можна за допомогою доопрацювання данних, використання більшої кількості негативних випадків, додавання експертного знання від спеціалістів банку та переугрупування вибірок.