НАЦІОНАЛЬНО ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

Моделі глибинного навчання для прогнозування дефолту позичальника на основі трансакційних та поведінкових даних

Виконав:
Студент 4-го курсу
Групи КА-51
Оніщук Олег Олегович

Науковий керівник:
кандидат технічних наук
Мілявський Юрій
Леонідович

Актуальність дослідження

Важливим елементом економіки багатьох розвинених країн є кредитування, вагому частку якого складають споживчі кредити.

Саме тому прогнозування дефолту позичальника є необхідним засобом оптимізації прийняття рішень щодо надання кредиту фізичним особам. А побудова ефективних моделей важливим інструментом для фінансово-кредитних установ.

Об'єкт дослідження

Прогнозування платоспроможності позичальників кредитно-фінансових установ.

Предмет дослідження

Моделі глибинного навчання у вигляді нейронних мереж.

Постановка задачі

1) Провести дослідження предметної області, а саме кредитування фізичних осіб, аналіз існуючих методів щодо прогнозування кредитоспроможності клієнтів фінансових установ;
2) Розглянути підходи до побудови моделей глибинного навчання у вигляді нейронних мереж для оцінки ймовірності дефолту позичальника;
3) Проаналізувати трансакційні, поведінкові, кредитні дані та розробити для них алгоритми побудови атрибутів;
4) Створити моделі нейронних мереж та дослідити ефективність кожної із них;
5) Визначити найкращу модель.

Кредитне оцінювання

В умовах зростання обсягів неповернення кредитів позичальниками особливої актуальності набуває питання оцінювання та управління банківськими кредитними ризиками.

При цьому банки, використовуючи світові передові технології та керуючись законодавчими і нормативно-правовими актами, впроваджують сучасні методики оцінювання та управління кредитними ризиками, розробляють методики оцінювання кредитних ризиків, які є найбільш прийнятними у їх діяльності і дають змогу максимально якісно оцінити ризик проведення кредитної операції.

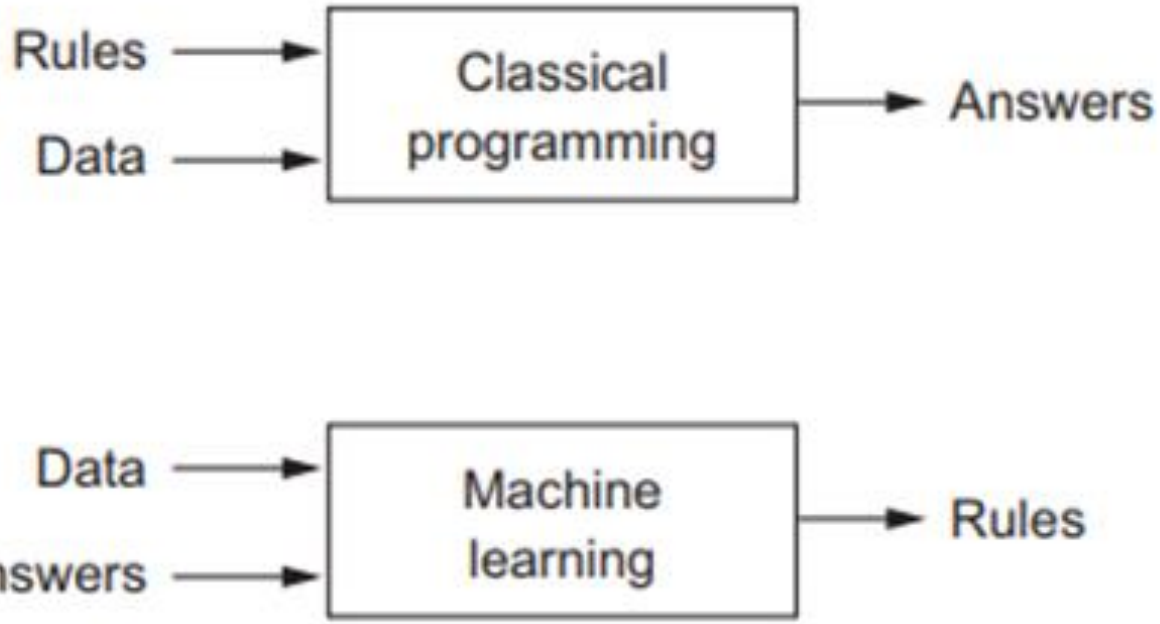
Методи кредитного оцінювання

Банківські установи можуть використовувати декілька методів, комбінуючи їх чи використовуючи різні методи для різних видів ризику. При цьому здійснюється аналіз як кількісної, так і якісної інформації про позичальника та ризику кредитної операції.

Підхід	Методологія
Методи судження	Застосування досвіду та розуміння аудитора у справі до рішення про продовження чи відмову у наданні кредиту
Експертні системи (наприклад, комітети з кредитування)	Застосування досвіду та розуміння аудитора у справі до рішення про продовження чи відмову у наданні кредиту
Аналітичні моделі	Використовується набір аналітичних методів, як правило, на кількісних даних, щоб отримати рішення

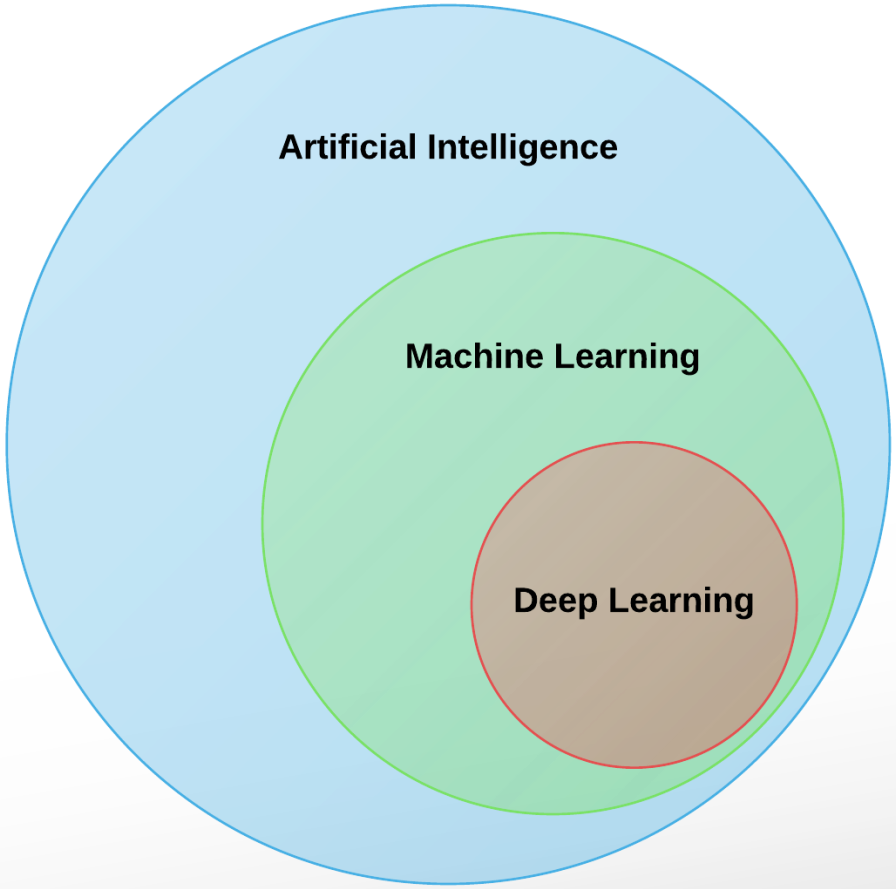
Підхід	Методологія
Статистичні моделі (наприклад, кредитна оцінка)	Використовується статистичний висновок для отримання відповідних відносин для прийняття рішень
Поведінкові моделі	Спостереження за поведінкою з часом, щоб отримати відповідні відносини для прийняття рішення
Моделі ринку	Покладаються на інформаційний зміст цін на фінансовому ринку як показники фінансової платоспроможності

Машинне навчання



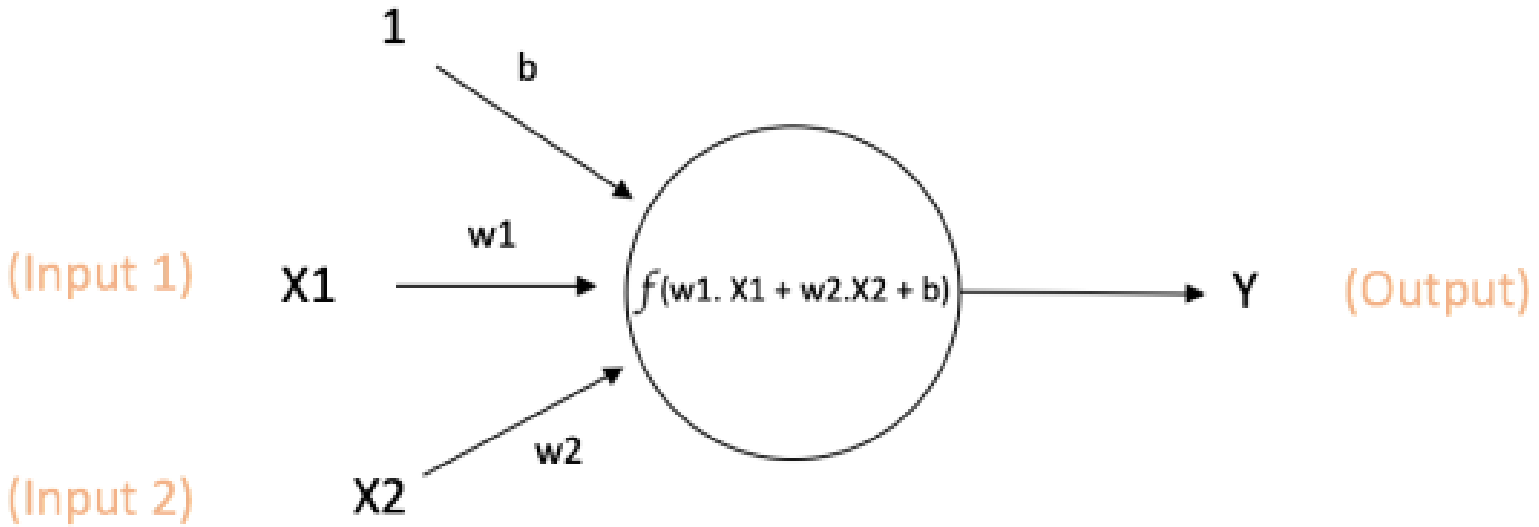
Систему машинного навчання тренують, а не програмують явно. Вона представлена багатьма прикладами, що мають пряме відношення до задачі, і знаходить статистичну структуру в цих прикладах, що в кінцевому підсумку дозволяє системі скласти правила автоматизації завдання.

Глибинне навчання



Глибинне навчання - це специфічне підполе машинного навчання: новий підхід до вивчення уявлень з даних, який робить акцент на вивченні більш глибоких взаємозвязків.

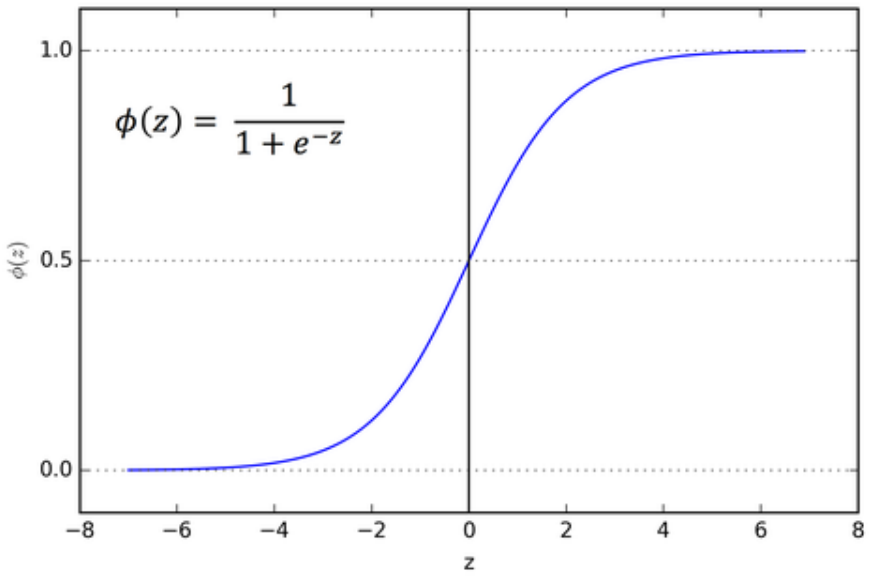
Нейрон



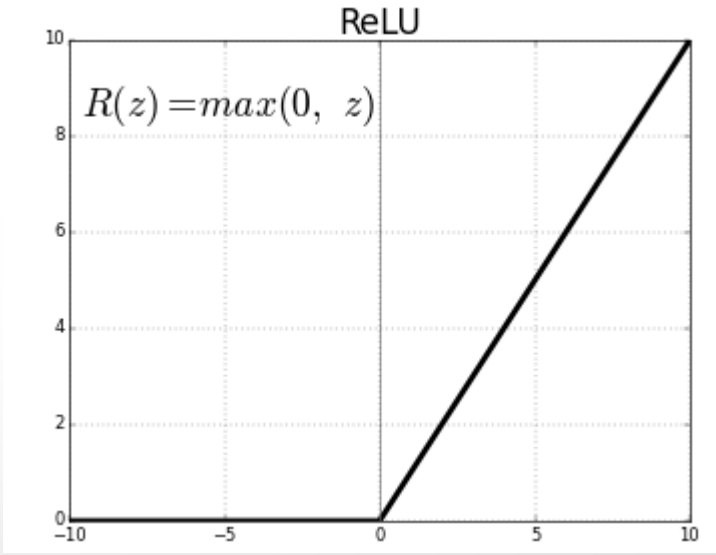
Output of neuron = Y= $f(w1.\ X1 + w2.X2 + b)$

Функції активації

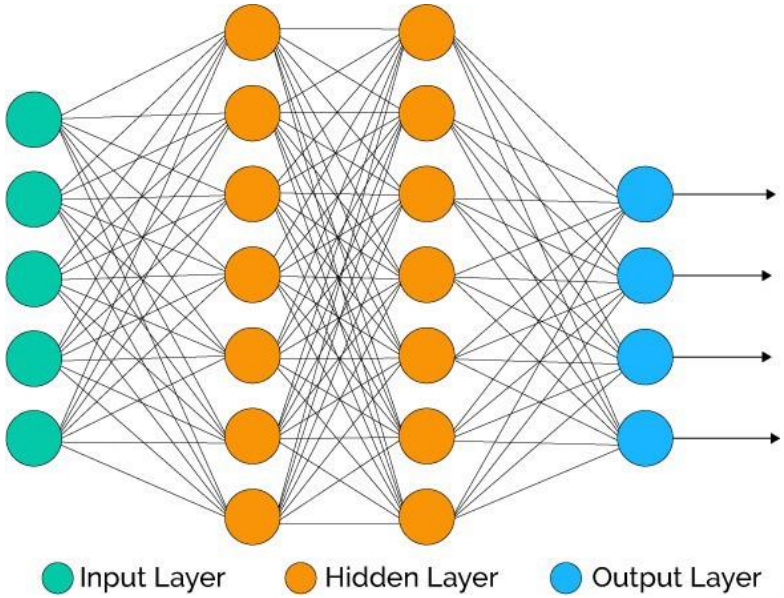
Сигмоїдна функція



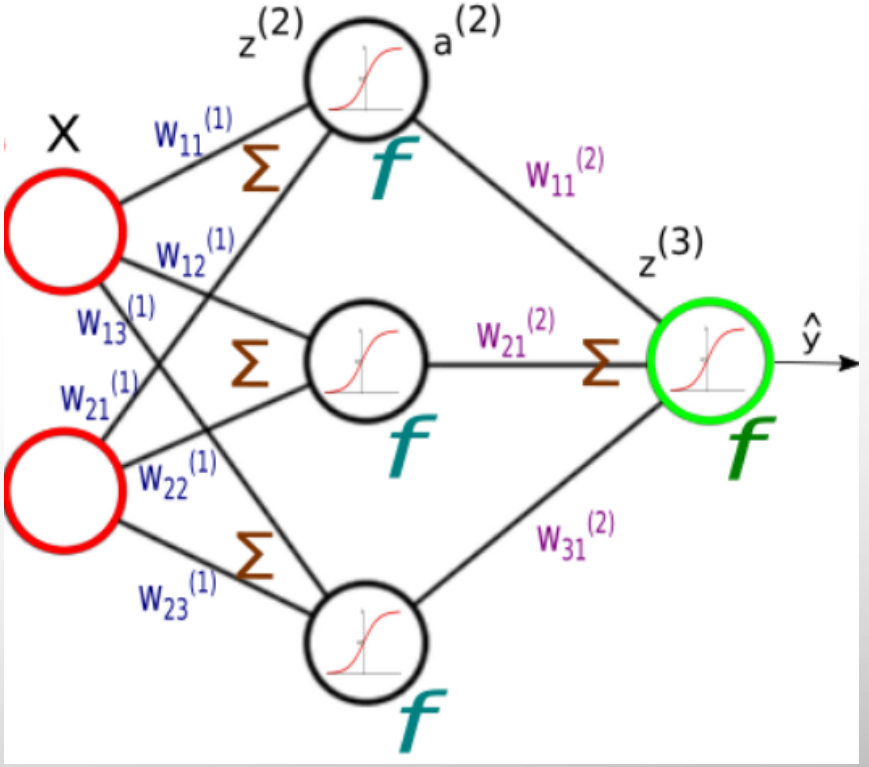
ReLU функція



Архітектура нейронних мереж



$$\hat{y} = f(W_{11}^{(2)} * a^{(2)} + \cdots)$$
$$a^{(2)} = f(W_{11}^{(1)} * x^{(1)} + \cdots)$$



Функція витрат

Loss (error) Function:

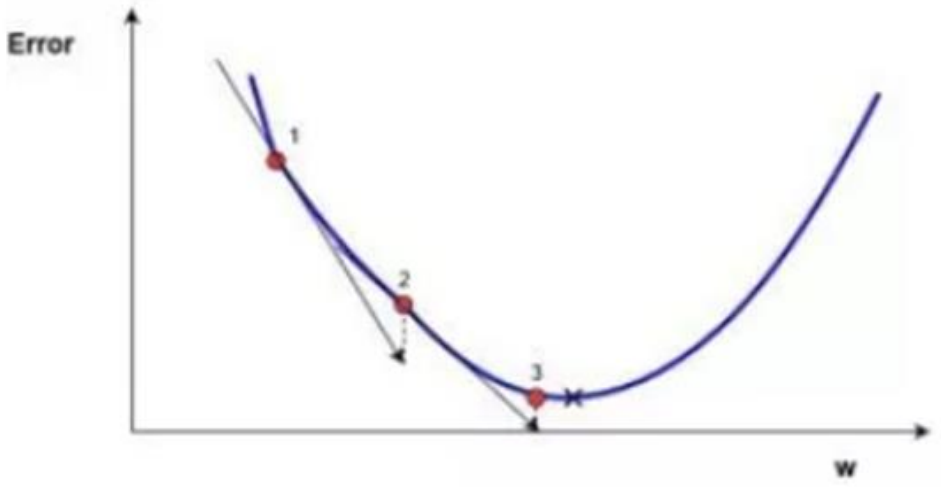
$$L(\hat{y}, y) = -(y * log\hat{y} + (1 - y) * \log(1 - \hat{y}))$$

Якщо $y = 1$: $L(\hat{y}, y) = -log\hat{y}$, хочемо, щоб $\hat{y} \uparrow$
Якщо $y = 0$: $L(\hat{y}, y) = -\log(1 - \hat{y})$, хочемо, щоб $\hat{y} \downarrow$

Cost Function:

$$J(W, b) = \frac{1}{m}\sum_{i=1}^{m} L\left(\hat{y}^i, y^i\right)$$

Градієнтний спуск



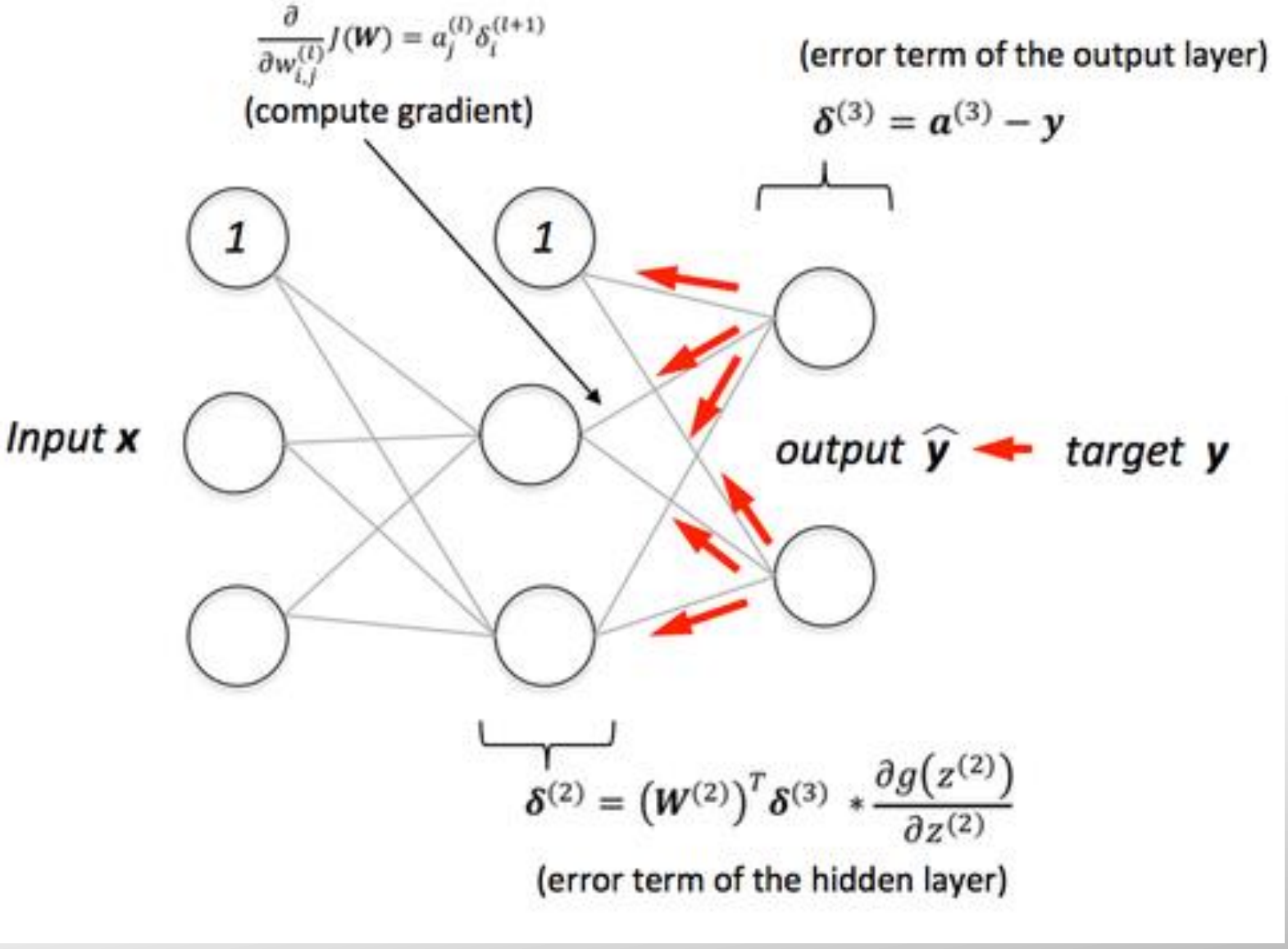
$$J(W,b) = \frac{1}{m}\sum_{z=0}^{m} J(W,b,x^{(z)},y^{(z)})$$

$$W = W - \alpha \nabla J(W,b)$$

$$w_{ij}^{(l)} = w_{ij}^{(l)} - \alpha \frac{\partial}{\partial w_{ij}^{(l)}} J(w,b)$$

$$b_{i}^{(l)} = b_{i}^{(l)} - \alpha \frac{\partial}{\partial b_{i}^{(l)}} J(w,b)$$

Алгоритм зворотного поширення



Трансакційні дані

Атрибут	Опис
advanceID	Унікальний ідентифікатор, який привязується до користувача-клієнта
txnDescription	Текстовий опис трансакції (наприклад, чи відбулася трансакція на автомат ATM – зняття готівки, отримання заробітної плати, платіжка по кредиту)
txnAmount	Числове значення трансакції
txnDate	Дата та час трансакції

Поведінкові дані

Атрибут	Опис
advanceID	Унікальний ідентифікатор, який привязується до користувача-клієнта
FieldId	Ідентифікатор поля, яке заповнює користувач. Може приймати 194 значення, для кожного з яких визначено своє унікальне поле
Value	Значення поля, яке вводилось
AvgSpeed	Середня швидкість друку
SdSpeed	Стандартне відхилення швидкості друку
FocusTimestamp	Час фокусування на полі
ChangeType	Тип події
PageId	Версія сайту (9 – десктопна версія, 121 - мобільна)

Кредитні дані

Атрибут	Опис
advanceID	Унікальний ідентифікатор, який привязується до користувача-клієнта
LoanNum	Числове значення номеру кредиту, який бере клієнт в цьому банку
ProductName	Назва виду кредитного продукту (тип кредитування)
AGDateCreated	Дата та час відкриття кредиту
Fails35NoPayIn90	Інформація про здійснення платежів протягом 90 днів (0 – здійснювались, 1 – не здійсювались)
LoanWriteOff	Інформація про повернення кредиту (0 – повернули, 1 – не повернули).

Обробка даних

Новий атрибут	Опис
Count	Загальна кількість трансакцій одного клієнта
mean	Середнє значення
std	Середньоквадратичне відхилення
Sum_tr_bigger0	Сума всіх позитивних транзакцій (значення > 0)
Num_tr_bigger0	Кількість всіх позитивних транзакцій
Sum_tr_smaller0	Сума негативних трансакцій
Sum_tr_smaller0	Кількість негативних трансакцій
…	…
Badtr_num / badtr_sum	Кількість та сума трансакцій в категорії «невдала трансакція»
…	…
Zay_num / zay_am	Кількість та сума трансакцій в категорії «служба зайнятості»

Нормалізація даних

$$X_{changed} = \frac{X - X_{min}}{X_{max} - X_{min}}$$

Стандартизація даних

$$X_{changed} = \frac{X - \mu}{\sigma}$$

Метрики

Середня абсолютна помилка:

$$\frac{1}{N}\sum_{j=1}^{N}|y_j - \hat{y}_j|.$$

Точність:

$$\frac{\text{К – сть правильних прогнозів}}{\text{Загальна кількість}}$$

Precision (точність) і recall (повнота):

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

Матриця помилок:

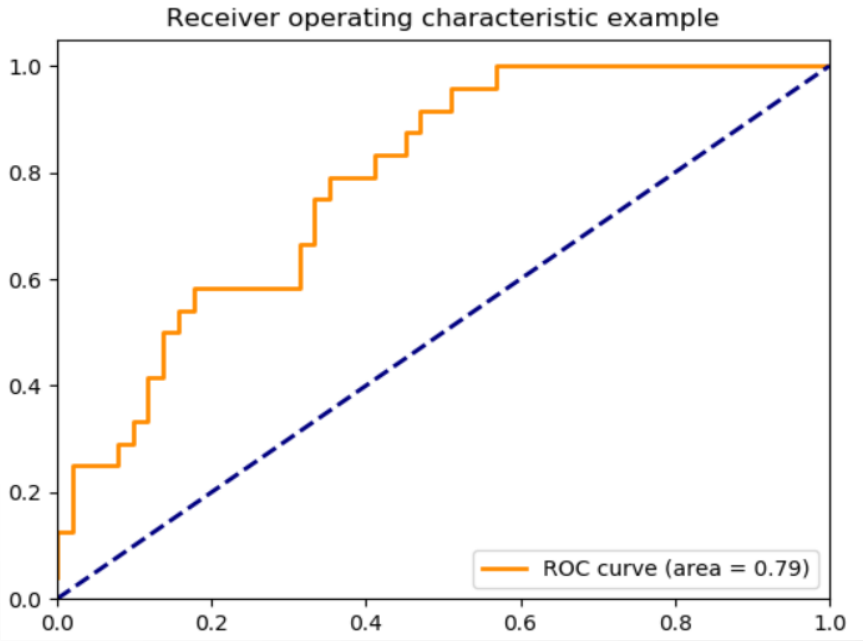
		Передбачуваний стан	
		1	0
Істинний стан	1	TP	FP
	0	FN	TN

Метрика ROC-AUC

$$\frac{\sum_{i=1}^{q}\sum_{j=1}^{q} I[y_i < y_j] I'[a_i < a_j]}{\sum_{i=1}^{q}\sum_{j=1}^{q} I[y_i < y_j]},$$

$$I'[a_i < a_j] = \begin{cases} 0, & a_i > a_j, \\ 0.5 & a_i = a_j, \\ 1, & a_i < a_j, \end{cases} \qquad I[y_i < y_j] = \begin{cases} 0, & y_i \geq y_j, \\ 1, & y_i < y_j, \end{cases}$$

де a_i – відповідь алгоритма на i-му об'єкті, y_i – його мітка(клас), q - число об'єктів в тесті.



Моделі нейронних мереж

Параметр	Модель 1	Модель 2	Модель 3
Кількість прихованих шарів	4	8	12
Кількість нейронів в прихованих шарах	128, 128, 64, 8	128, 128, …, 128, 64, 8	128, 128, 128, ..., 128, 64, 8
Активація	ReLu	ReLu	Softmax
Функція втрат	binary_cross	binary_cross	binary_cross
Оптимізація	SGD	SGD	SGD
Epoch	20	20	20
Batch size	128	128	128

Результати прогнозування

Дані	ROC_AUC		
	Модель 1	Модель 2	Модель 3
Трансакційні 1-ий варіант	0.61976	0.64256	0.63548
Трансакційні 2-ий варіант	0.735464	0.752354	0.7412356
Поведінкові	0.61894	0.63789	0.623265

Порівняння результатів

Як можна побачити з попередньої таблиці, найкращий результат дала модель №2 на трансакційних даних. Ця модель є найкращою з трьох, адже в моделі 1 занадто мало шарів і нейронів, а в моделі 3 навпаки – занадто багато і відбувається перенавчання.

ROC-AUC = 0.752354

Висновки

1) Проведено дослідження предметної області, а саме кредитування фізичних осіб, аналіз існуючих методів до прогнозування кредитоспроможності клієнтів фінансових установ;
2) Розглянуто підходи до побудови моделей глибинного навчання у вигляді нейронних мереж для оцінки ймовірності дефолту позичальника;
3) Проаналізовано трансакційні, поведінкові, кредитні дані та розроблено для них алгоритми побудови атрибутів;
4) Створено моделі нейронних мереж та досліджено ефективність кожної із них;
5) Визначено найкращу модель.

Дякую за увагу!