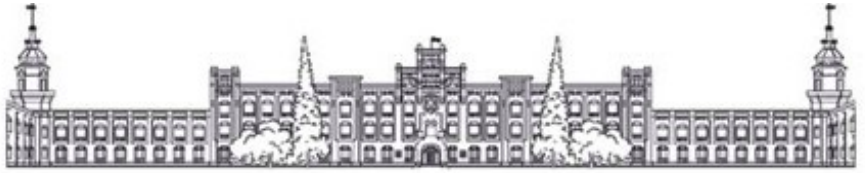
НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМЕНІ
ІГОРЯ СІКОРСЬКОГО»
ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ
КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

МОДЕЛІ КРЕДИТУВАННЯ ФІЗИЧНИХ ОСІБ ЗА ДОПОМОГОЮ НЕЙРОННИХ МЕРЕЖ

Виконавець роботи:
студент 4 курсу
групи КА-43
Куриленко Олександр Михайлович

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

19 червня 2018 року



Мета дослідження:

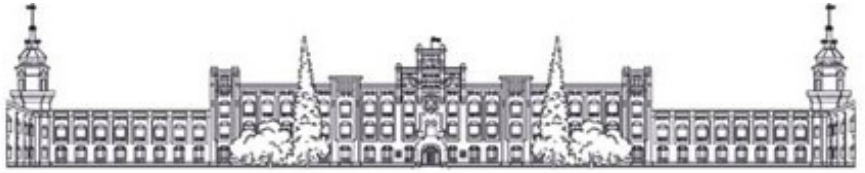
виконання процесу кредитного оцінювання шляхом оцінки кредитного ризику для фізичних осіб.

Об’єкт дослідження:

набір історичних даних, взятих з Kaggle-змагання ‘Give Me Some Credit’ щодо наданих фінансовою установою-кредитором кредитів фізичним особам.

Предмет дослідження:

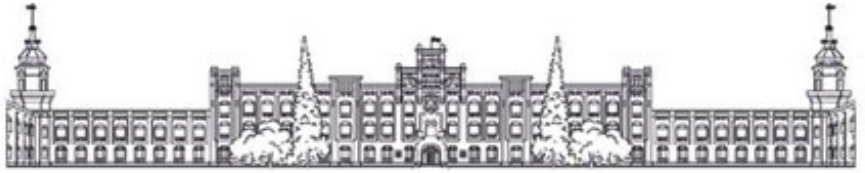
нейронні мережі, випадковий ліс, метод k найближчих сусідів



Актуальність дослідження

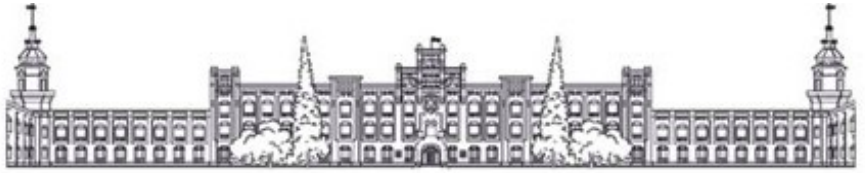
Кредитне оцінювання є засобом оптимізації процесу прийняття рішень щодо надання кредитних коштів фізичним особам в рамках споживчих кредитів.

Тож дане дослідження, тобто побудова адекватних моделей моделей кредитного оцінювання, є необхідним для фінансових установ-кредиторів.



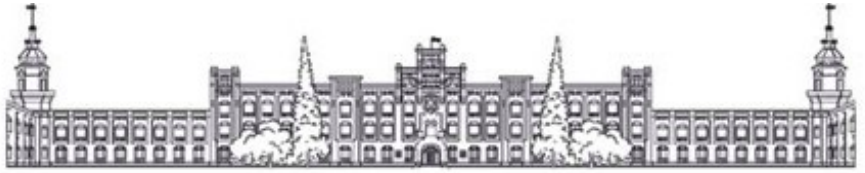
Постановка задачі

- провести дослідження існуючих підходів до вирішення задачі кредитного оцінювання, розглянути історію розвитку даних аналітичних підходів в області кредитних ризиків;
- розглянути підходи до побудови моделей оцінки кредитного ризику за допомогою нейронних мереж, а також випадкового лісу та методу k найближчих сусідів, та розглянути їх алгоритми роботи;
- проаналізувати вхідні дані, провести їх попередню обробку, а також розробити архітектуру моделей оцінки кредитного ризику, що після визначення порогових значень прийняття рішень можуть використовуватися в якості частини СППР або в якості СППР;
- розробити модель для оцінки кредитного ризику за допомогою нейронних мереж та провести порівняння результатів цієї моделі з результатами, отриманими іншими методами, що розглядаються;
- для найкращих моделей визначити бінарні класифікатори.



Базові поняття процесу кредитування

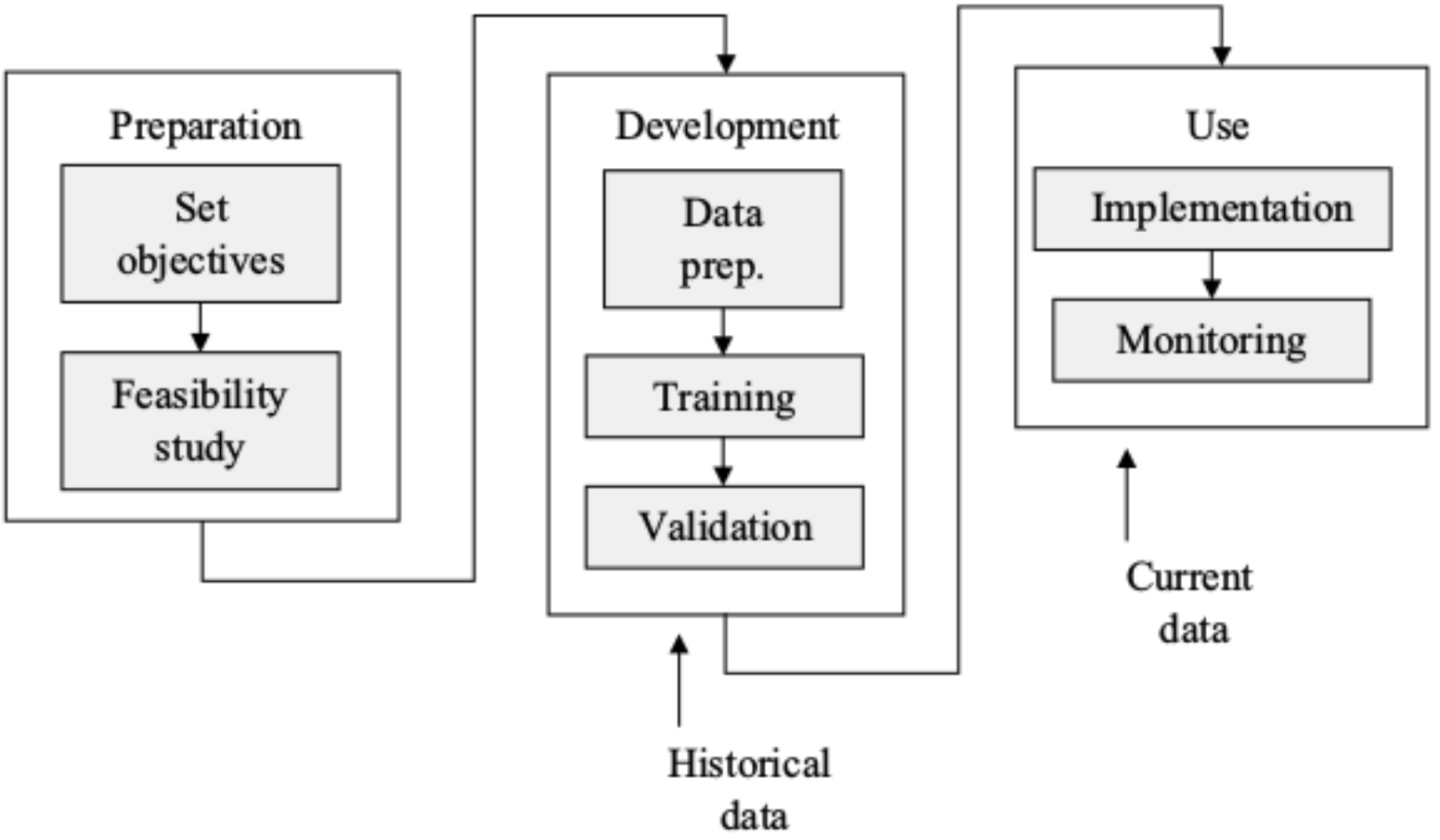
- **Кредитоспроможність** – бажання та здатність позичальника погасити борг.
- **Кредитний ризик** – потенційний фінансовий вплив будь якої імовірної або реальної зміни в кредитоспроможності позичальника; тобто ризик виникнення дефолту дебітора (позичальника).

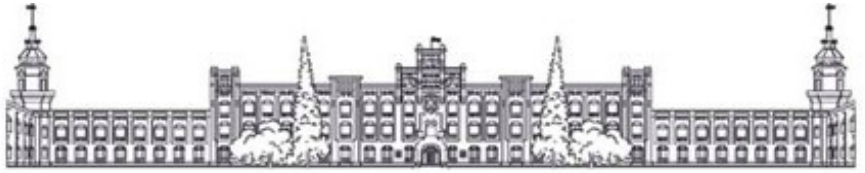


Кредитне оцінювання

Кредитне оцінювання – це використання статистичних моделей для перетворення відповідних даних в чисельні міри, що є основою для рішень щодо видачі або невидачі кредиту фізичній особі на певних умовах.

Процес кредитного оцінювання

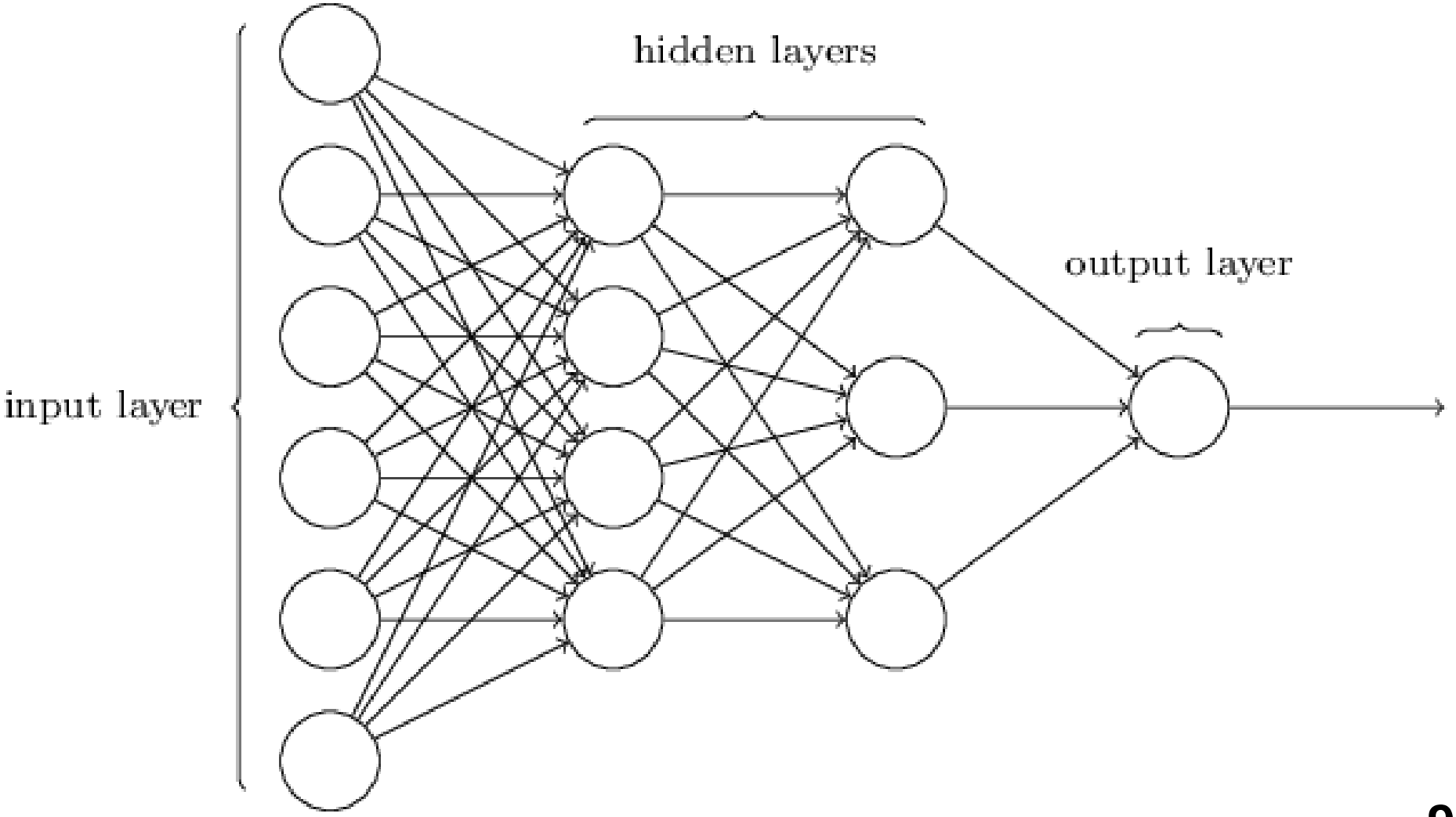


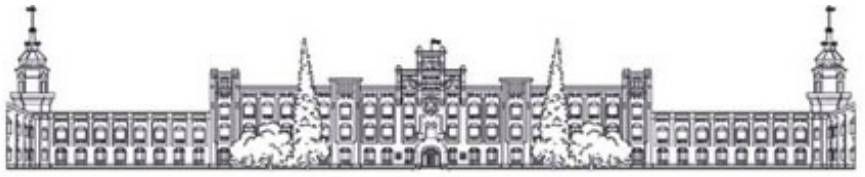


Математичні методи для побудови моделей оцінки кредитного ризику

- Нейронні мережі
- Випадковий ліс
- Метод k найближчих сусідів

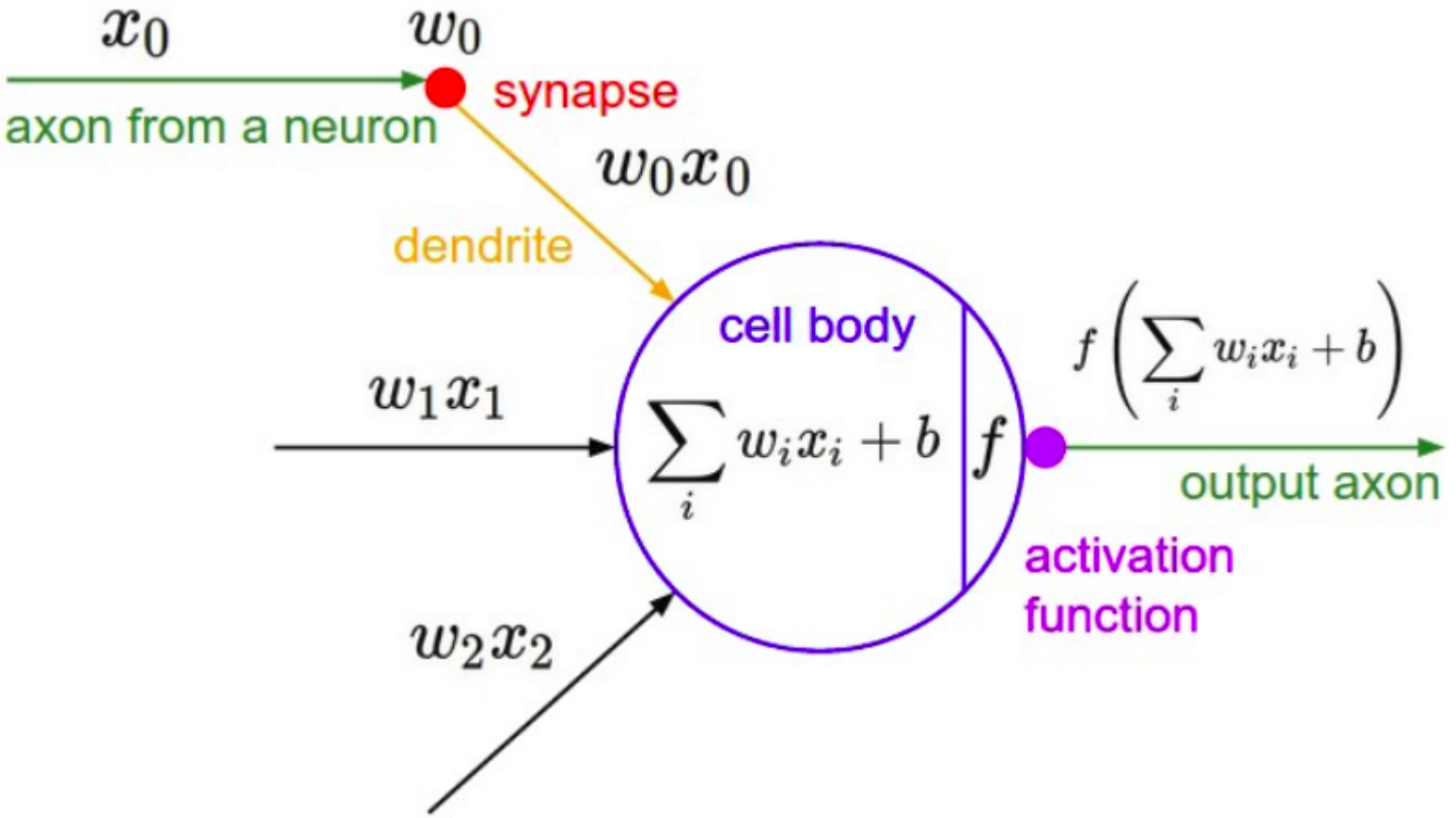
Нейронні мережі

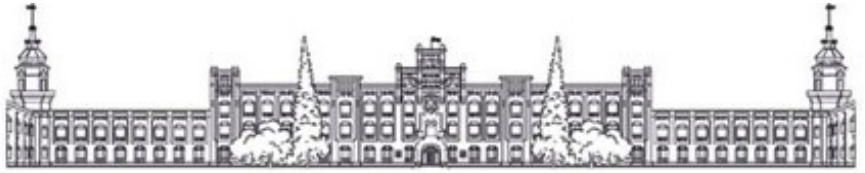




Нейронні мережі

$$output = \begin{cases} 0, & w \cdot x + b \leq 0, \\ 1, & w \cdot x + b > 0, \end{cases}$$ if there's no activation function applied.





Нейронні мережі

- Активаційні функції.
- Функція втрат $J(w,b), J(w,b) \to min.$

Для бінарної класифікації обирається a) логарифмічна функція правдоподібності або b) крос-ентропія (для 2-х виходів):

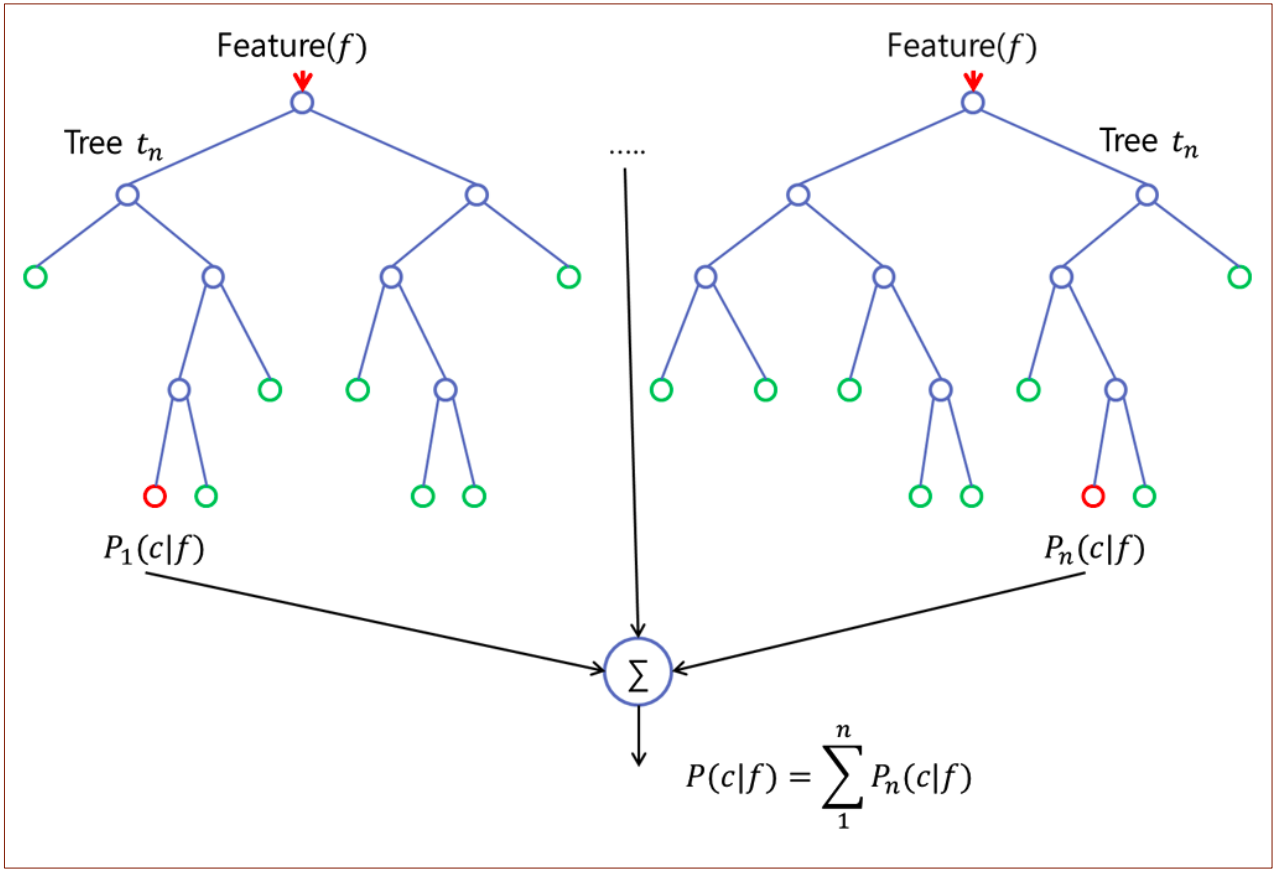
a) $J(w,b) = -\frac{1}{n} \cdot \sum_x ln(a_y^L),$ b) $J(w,b) = -\frac{1}{n} \cdot \sum_x \sum_j [y_j \cdot ln(a_j^L) + (1-y_j) \cdot ln(1-a_j^L)].$

- Оптимізований метод стохастичного градієнтного спуску для $J(w,b) \to min$ (оптимізації AdaGrad, AdaDelta, Adam тощо).
- Метод зворотнього поширення помилки для обчислення градієнта на кожному кроці методу стохастичного градієнтного спуску.

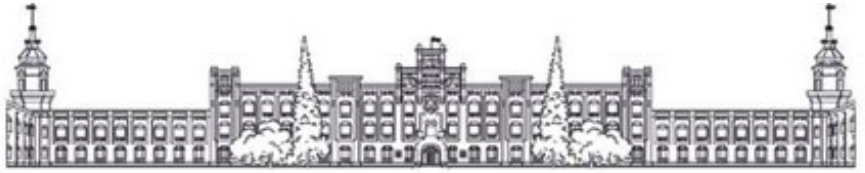
Випадковий ліс

Підбір кількості дерев рішень L та навіть значення m (к-сть змінних, що випадково відбираються на кожному поділі) в випадковому лісі проводиться за критерієм out-of-bag:

кількість помилок на вхідних екземплярах x_i, якщо не враховувати 'голоси' дерев, на яких x_i був навчальним прикладом (тобто належав підвибірці X_l, l=1,2,…,L)



$$out\text{-}of\text{-}bag(a) = \sum_{i=1}^{N}[sign(\sum_{l=1}^{L}[x_i \notin X_l]\ b_l(x_i)) \neq y_i] \rightarrow min.$$



Метод k найближчих сусідів

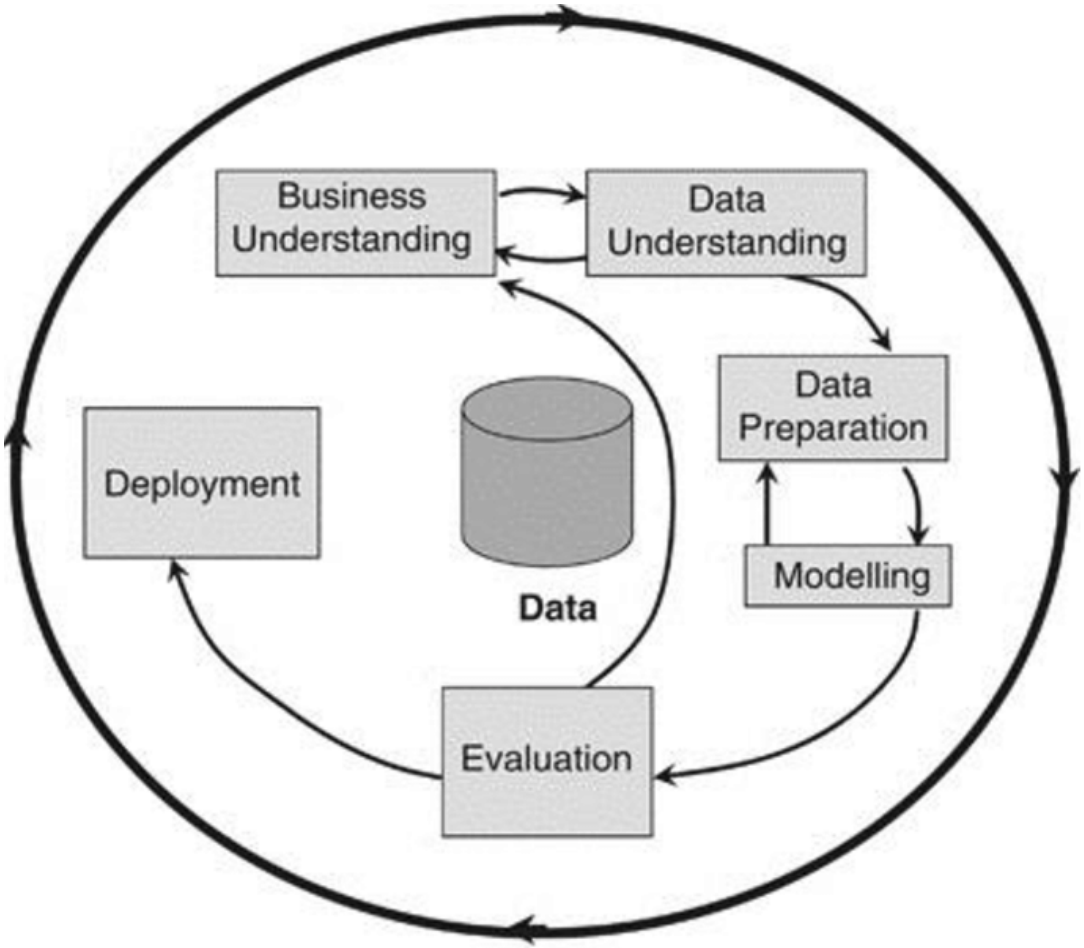
- Метод k найближчих сусідів часто використовується завдяки простоті інтерпретацій його результатів та малому часу, необхідному для обчислень.
- В методі k найближчих сусідів k – кількість найближчих за обраною метрикою (відстанню) сусідів певної точки, на класи яких він спирається при визначенні її класу.

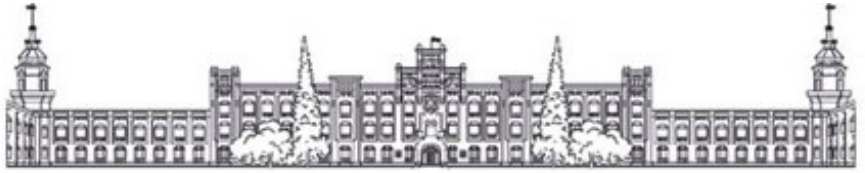
Приклади відстаней для простору розмірності N:

- Евклідова (кількісні): $\mathrm{d}_{ik} = \left(\sum_{j=1}^{N} (x_{ij} - x_{kj})^2\right)^{\frac{1}{2}}$
- Манхетенська (кількісні): $\mathrm{d}_{ik}^{(l)} = \sum_{j=1}^{N} |x_{ij} - x_{kj}|$
- Журавльова (змішані): $\mathrm{d}_{ik} = \sum_{j=1}^{N} I_{ik}^{j},\ I_{ik}^{j} = \begin{cases} 1, & |x_{ij} - x_{kj}| < \varepsilon \\ 0, & \text{otherwise} \end{cases}$

Методологія дослідження

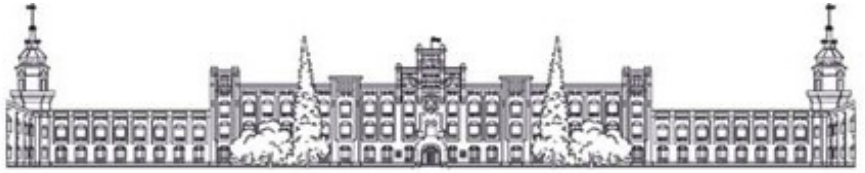
Для виконання поставлених завдань ми дотримувалися CRISP-DM (Cross-industry standard process for data mining) методології дослідження.





Модель життєвого циклу дослідження

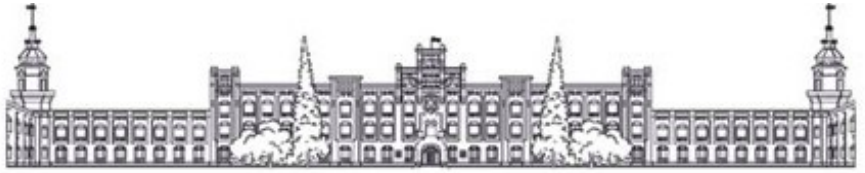
- ознайомлення з предметною областю та вивчення даних;
- підготовка даних до моделювання
- побудова нейронної мережі, випадкового лісу та застосування алгоритму k найближчих сусідів;
- оцінка результатів;
- впровадження.



Вивчення даних

- Початкова вибірка даних з Kaggle-змагання ‘Give Me Some Credit’ складалася з 12 змінних та 150000 рядків.
- Через неважливість змінну Id було відкинуто, а інші змінні перейменовано для зручності розуміння. В результаті отримали 11 змінних:

Target, BalanceToCredit, Age, Num1, DebtToIncome, Income, NumLoans, Num3, NumRealEstateLoans, Num2, NumDependents.

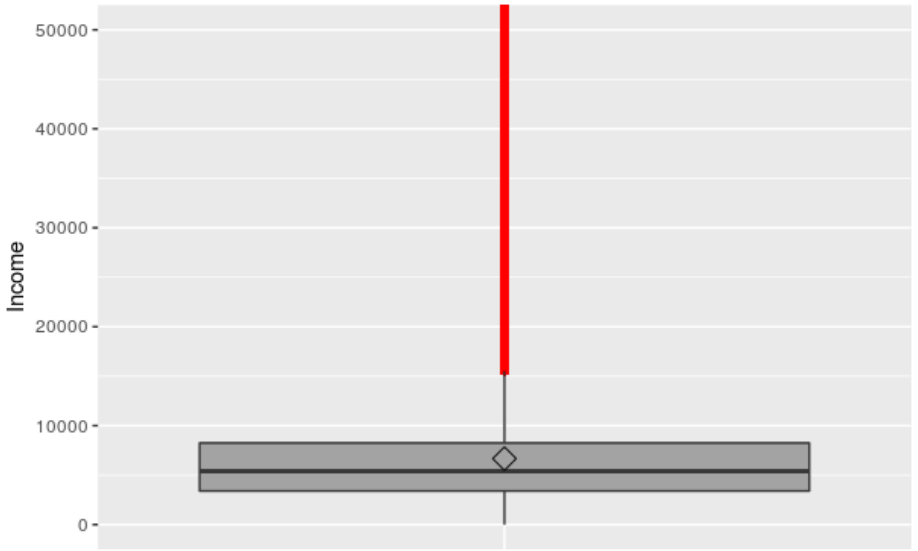


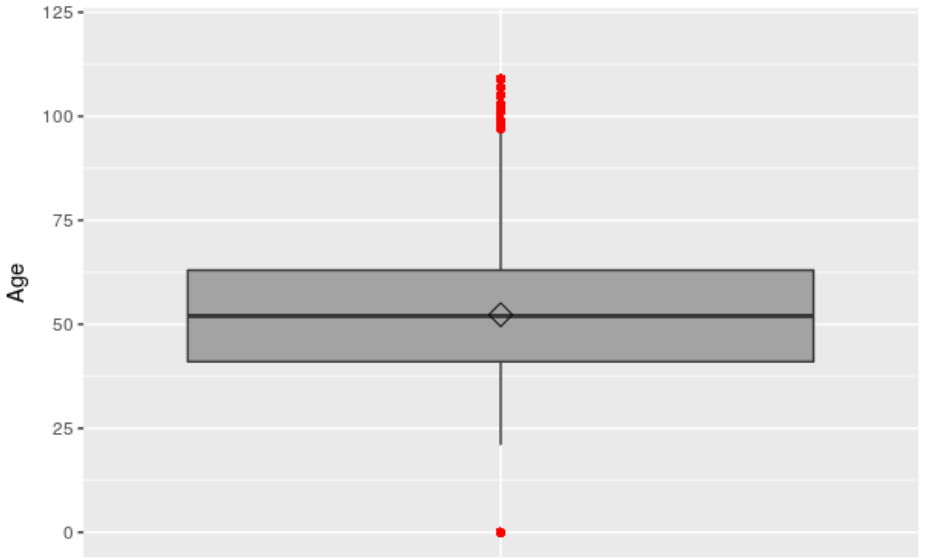
Вивчення даних

- Target – чи зазнає особа прострочки в 90 днів або гірше впродовж наступних 2 років;
- BalanceToCredit – загальний баланс на кредитних картках та персональних кредитних лініях за винятком заборгованостей на нерухомість та заборгованостей в розстрочку, поділений на суму кредитних лімітів;
- Age – вік позичальника в роках;
- Num1 – кількість разів, коли позичальник за останні 2 роки мав прострочку платежу в 30-59 днів, але не гірше;
- DebtToIncome – місячні виплати за боргами, аліменти, витрати на життя, поділені на місячний валовий дохід;
- Income – місячний дохід;
- NumLoans – кількість відкритих позик та кредитних ліній;
- Num3 – кількість разів, коли позичальник за останні 2 роки мав прострочку платежу в 90 днів або більше;
- NumRealEstateLoans – кількість іпотечних кредитів та кредитів на нерухомість включаючи кредитні лінії під заставу житлової нерухомості;
- Num2 – кількість разів, коли позичальник за останні 2 роки мав прострочку платежу в 60-89 днів, але не гірше;
- NumDependents – кількість утриманців в сім'ї за винятком позичальника.

Обробка даних

Для обробки даних кожна змінна досліджувалася поокремо. Для цього будувалися графіки ‘ящик з вусами’ для кожної зі змінних, а також графіки залежності кожної зі змінних від індекса в наборі даних.

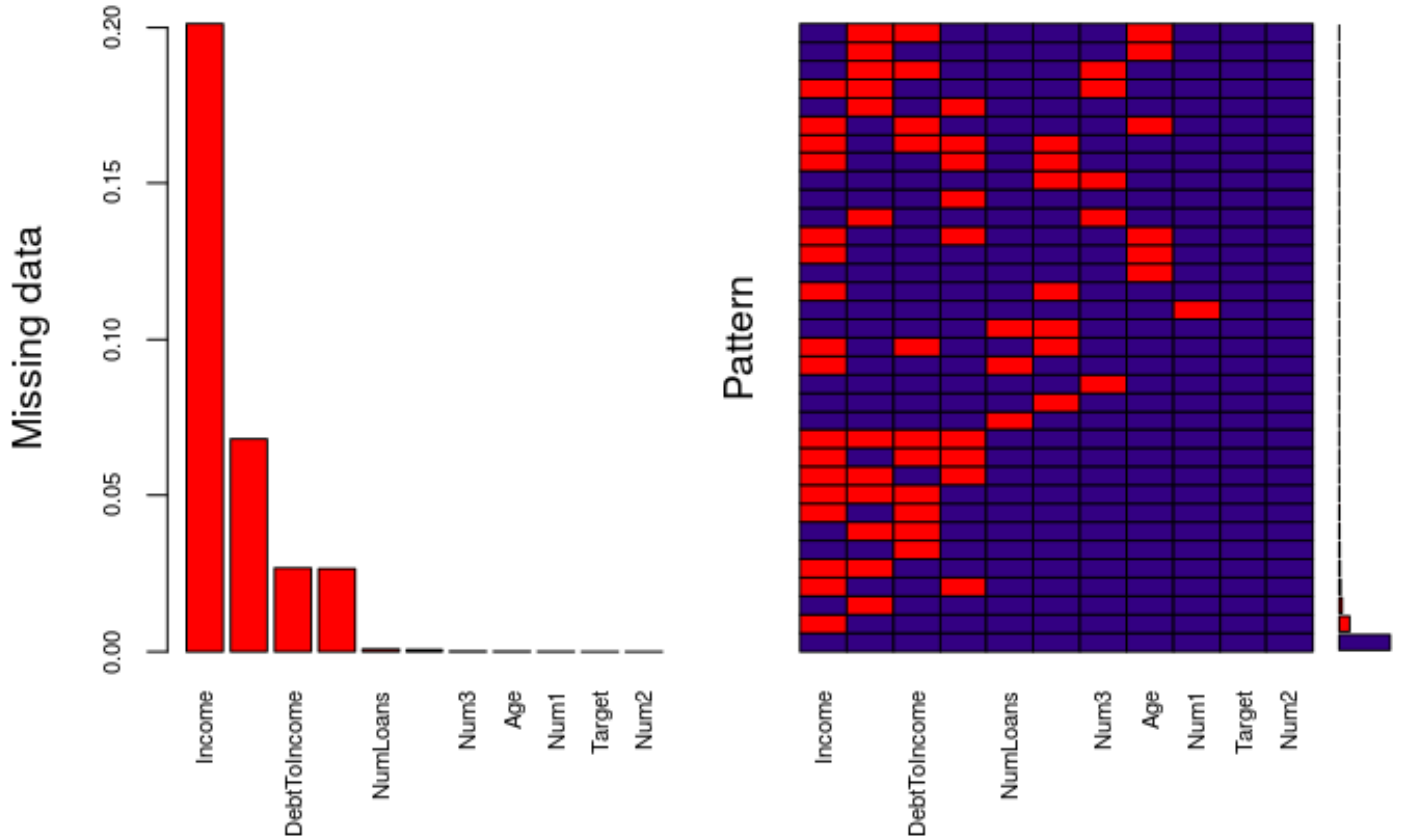


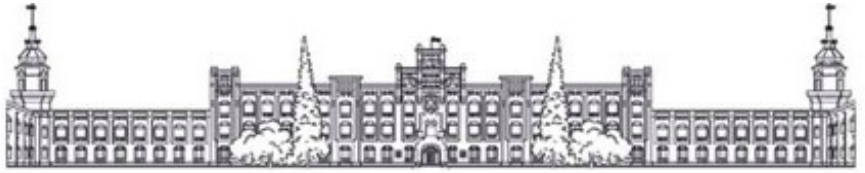


Графіки ‘ящик з вусами’ для змінних Income та Age

Підготовка даних до моделювання

Під час обробки даних визначалися можливі викиди та замінювалися на порожні значення – NA. В результаті було отримано набір даних з великою кількістю пропущених значень, які замінювалися предиктивним методом.





Нормалізація даних

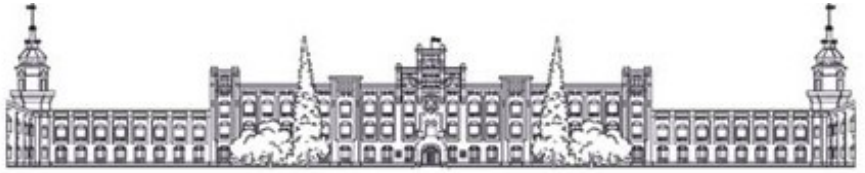
При нормалізації даних для методу k найближчих сусідів застосували наступні перетворення:

- Йео-Джонсона до змінних Age, Income, NumLoans, NumRealEstateLoans;

  λ варіюється і обирається найкраще!

$$\psi(\lambda, y) = \begin{cases} \frac{(y+1)^{\lambda}-1}{\lambda}, & \lambda \neq 0, y \geq 0; \\ log(y+1), & \lambda = 0, y \geq 0; \\ -\frac{(-y+1)^{2-\lambda}-1}{2-\lambda}, & \lambda \neq 2, y < 0; \\ -log(-y+1), & \lambda = 2, y < 0. \end{cases}$$

- $y = log_{10}(0.4 + x)$ до змінної BalanceToCredit;
- $y = log_{10}(x)$ до змінної DebtToIncome;
- $z = \frac{x - \mu}{\sigma}$ до змінних NumDependents, Num1, Num2, Num3.



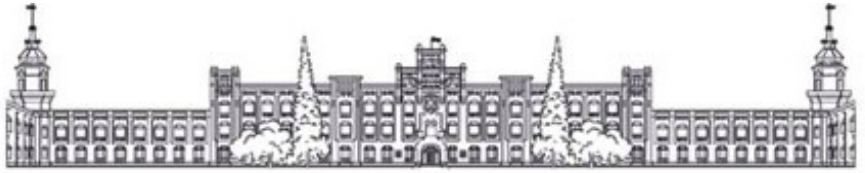
Матриця помилок

		Predicted Class	
		No	Yes
Observed Class	No	TN	FP
	Yes	FN	TP

TN	True Negative
FP	False Positive
FN	False Negative
TP	True Positive

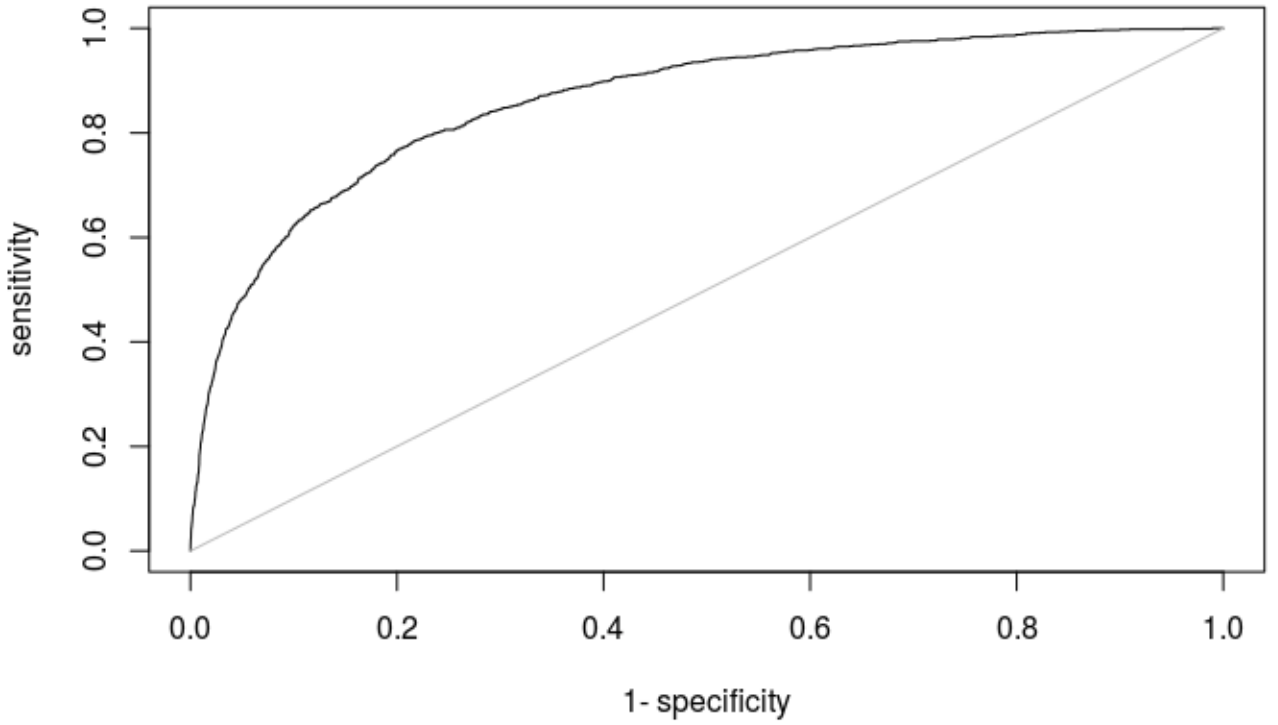
Model Performance

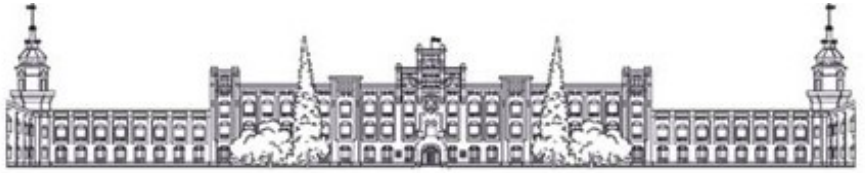
Accuracy	= (TN+TP)/(TN+FP+FN+TP)
Precision	= TP/(FP+TP)
Sensitivity	= TP/(TP+FN)
Specificity	= TN/(TN+FP)



Метрика для оцінки моделей

В якості метрики для оцінки моделей було обрано AUC – площу під ROC, адже ця метрика не вимагає визначення порогового значення прийняття рішень.





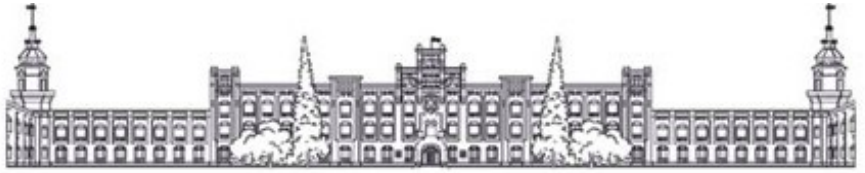
Побудова нейронної мережі

- Для методу стохастичного градієнтного спуску застосовуємо оптимізацію Adam:

$$m_{t,i} = \beta_1 \cdot m_{t-1,i} + (1-\beta_1) \cdot g_{t,i};$$
$$v_{t,i} = \beta_2 \cdot v_{t-1,i} + (1-\beta_2) \cdot g_{t,i}^2; \;\; g_{t,i} = \nabla_\Theta J(\Theta_{t,i});$$
$$\Theta_{t+1,i} = \Theta_{t,i} - \frac{\eta}{\sqrt{\hat{v}_{t,i} + \epsilon}} \cdot \hat{m}_{t,i}, \;\; \Theta - \text{network parameter}.$$

- Швидкість навчання визначаємо наступним чином:

$$\eta_1 = const;$$
$$\eta_{t+1} = \eta_t \cdot \frac{1}{1 + t \cdot DecayRate}.$$

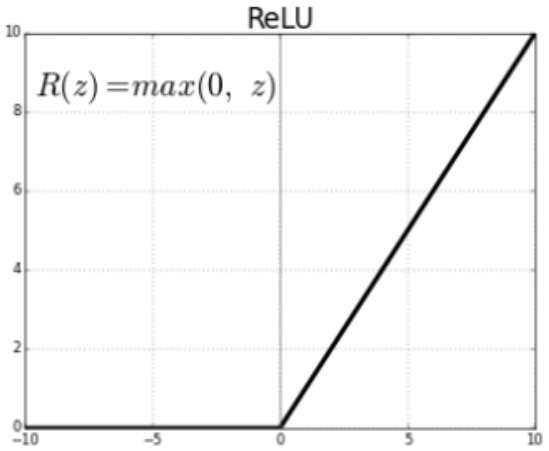


Adam:

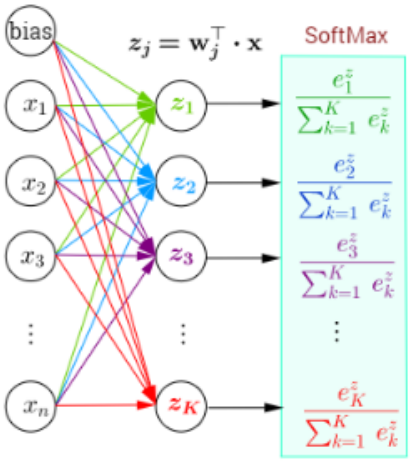
$$\begin{cases} \beta_1 = 0.9; \\ \beta_2 = 0.999; \\ \epsilon = 10^{-7}. \end{cases}$$

η_1	0.00002
Кількість епох навчання	180
$DecayRate$	0.0000014
Метод оптимізації стохастичного градієнтного спуску	Adam
$HiddenLayersN$	3
$NodesN$	350
Активаційна функція в вхідному та прихованих шарах	ReLU
Активаційна функція в вихідному шарі	Softmax
$ValidationSplit$	0.33
Розміри міні-батчів m та $EarlyStoppingEpochs$ ($m\_EarlyStoppingEpochs$)	710_3, 640_3, 640_3, 840_3, 900_3, 900_3, 900_3, 900_3, 1000_2, 10000_2, 10000_2, 10000_3, 10000_3, 12000_4, 14000_4, 10000_4, 15000_4, 16000_4, 14000_4, 14000_4, 14000_4

Результати роботи нейронної мережі



(а)

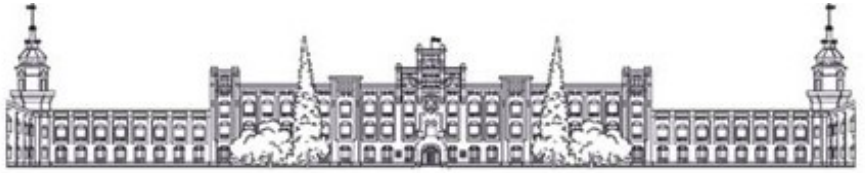


(б)

Активаційні функції: (а) ReLU; (б) Softmax

На тестовій вибірці було отримано результат AUC = 0.85176.

Функцію втрат на тестовій вибірці було мінімізовано до 0.17836.



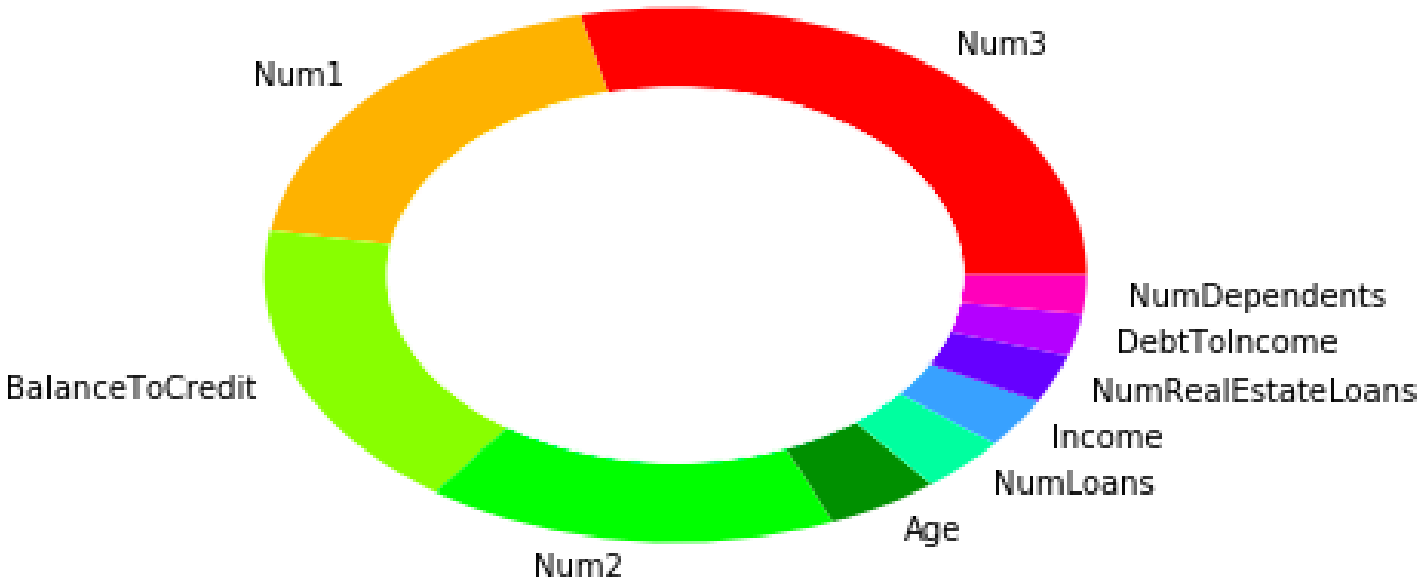
Випадковий ліс

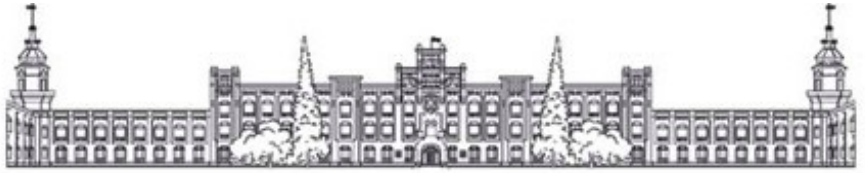
- Для випадкового лісу найкращою виявилася модель з наступними параметрами:

  $L = 2000;$
  $m = 2;$
  $min.node.size = 60.$

- На тестовій вибірці було отримано результат AUC = 0.86204.

Важливість змінних для оцінки кредитного ризику



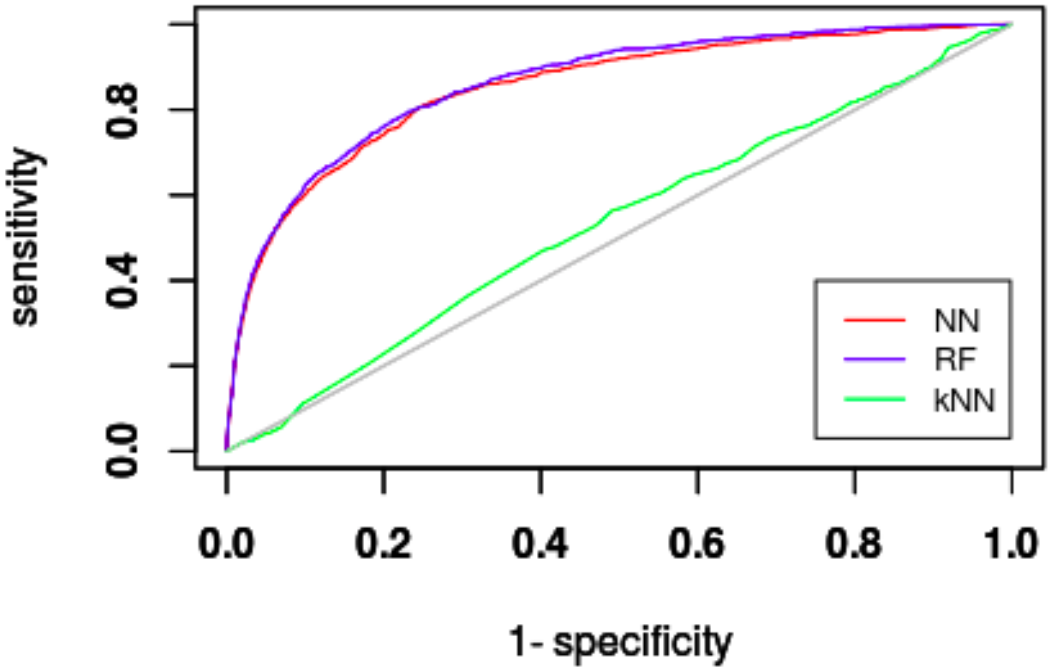


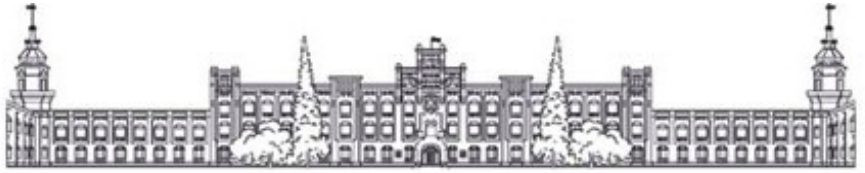
Метод k найближчих сусідів

Незважаючи на простоту інтерпретації, на тестовій вибірці показав максимальне значення AUC = 0.53288 при k = 150.

Порівняння результатів

- Нейронна мережа: AUC = 0.85176.
- Випадковий ліс: AUC = 0.86204.
- Метод k найближчих сусідів: AUC = 0.53288.





Впровадження

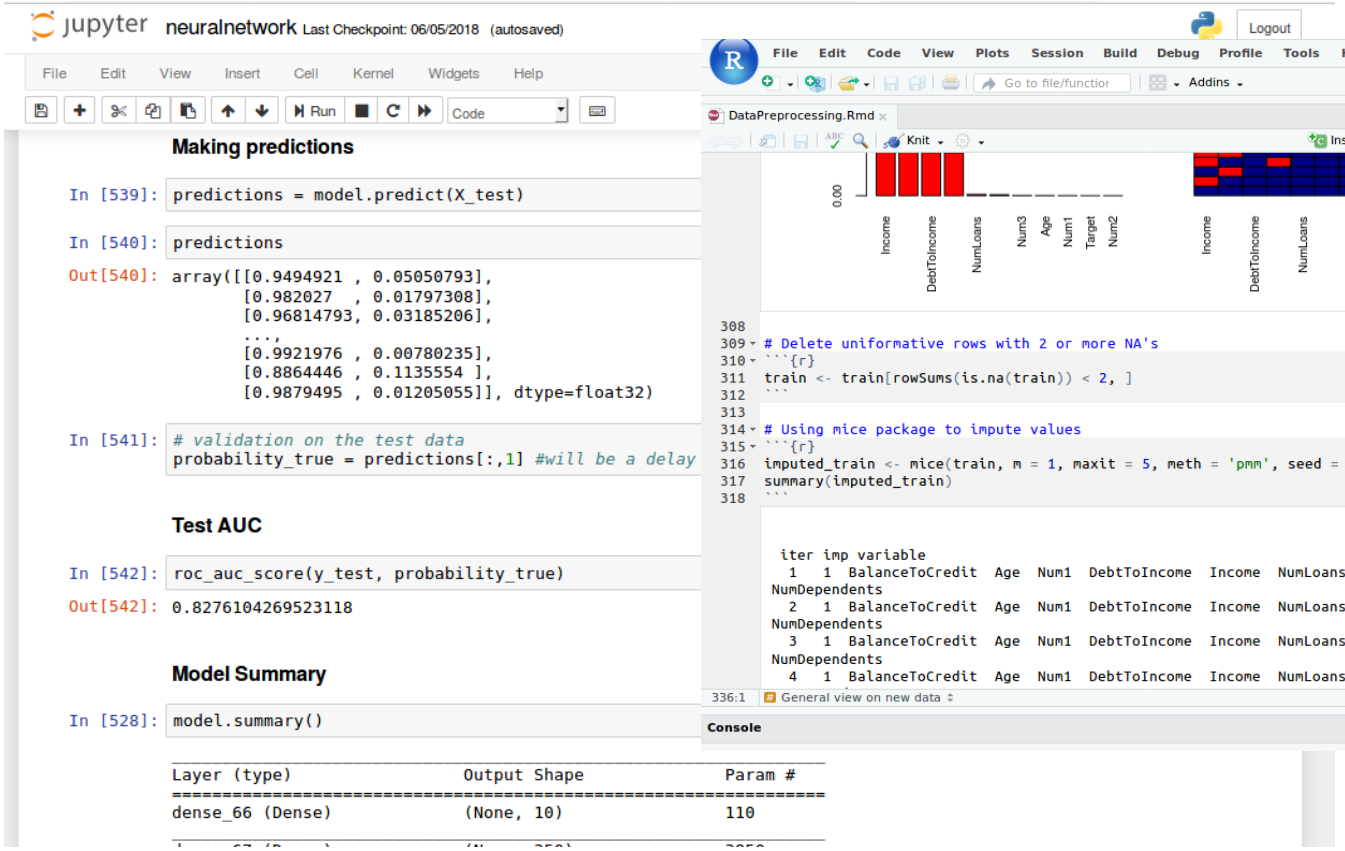
В якості впровадження перетворюємо моделі – випадковий ліс та нейронну мережу – в бінарні класифікатори, що можуть бути застосовані як частина СППР або в якості СППР.

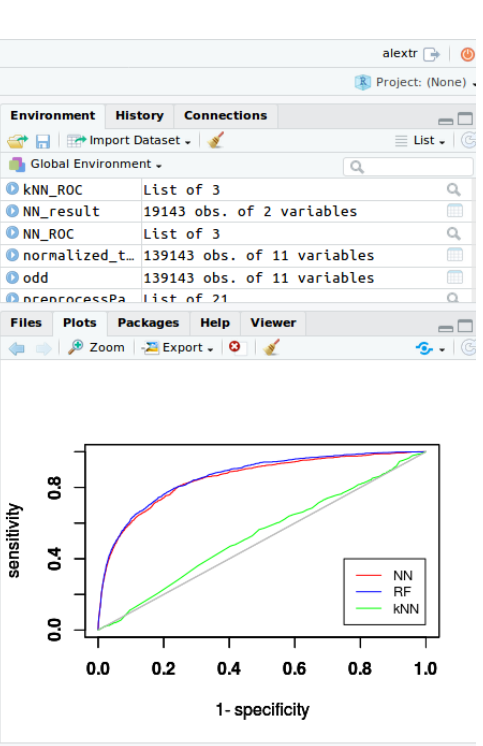
Для цього визначаємо порогове значення прийняття рішень $cutoff$.

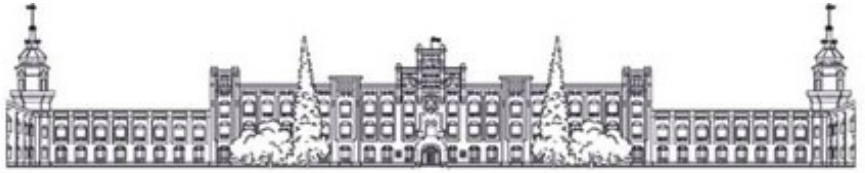
Для нейронної мережі: $cutoff = 0.057067$, $specificity \approx 0.773$, $sensitivity \approx 0.774$.
Для випадкового лісу: $cutoff = 0.060575$, $specificity \approx 0.781$, $sensitivity \approx 0.781$.

Програмний продукт

В результаті було отримано ПП, що складається з 6 окремих модулів на мовах R і Python: обробка даних, нейронна мережа, дослідження нейронної мережі та визначення порогового значення прийняття рішень, випадковий ліс, метод k найближчих сусідів, побудова ROC-кривих.

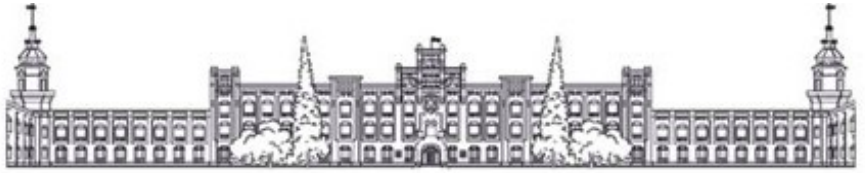






Висновки

- проведено дослідження існуючих підходів до вирішення задачі кредитного оцінювання, розглянуто історію розвитку даних аналітичних підходів в області кредитних ризиків;
- розглянуто підходи до побудови моделей оцінки кредитного ризику за допомогою нейронних мереж, а також випадкового лісу та методу k найближчих сусідів, а також розглянуто їх алгоритми роботи;
- проаналізовано вхідні дані, проведено їх попередню обробку, а також розроблено архітектуру моделей оцінки кредитного ризику, що після визначення порогових значень прийняття рішень можуть використовуватися в якості частини СППР або в якості СППР;
- розроблено модель для оцінки кредитного ризику за допомогою нейронних мереж та проведено порівняння результатів цієї моделі з результатами, отриманими іншими методами, що розглядалися;
- для найкращих моделей визначено бінарні класифікатори.



Перспективи щодо подальших досліджень

- накопичення більшої кількості історичних даних та вдосконалення наявних моделей, адже на більших кількостях даних, наприклад, нейронна мережа даватиме кращий результат;
- виконання процесу точного визначення порогового значення прийняття рішень для отримання з побудованих моделей бінарних класифікаторів в залежності від конкретної бізнес-моделі фінансової установи-кредитора.

Публікації

Опубліковано тези доповідей на XX міжнародній науково-технічній конференції SAIT 2018:

‘Credit scoring models development’

Kurylenko O. M.

ДЯКУЮ ЗА УВАГУ!