МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАВЧАЛЬНО-НАУКОВИЙ КОМПЛЕКС

"ІНСТИТУТ ПРИКЛАДНОГО СИСТЕМНОГО АНАЛІЗУ"

НАЦІОНАЛЬНОГО ТЕХНІЧНОГО УНІВЕРСИТЕТУ УКРАЇНИ

"КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ ІМ.І.СІКОРСЬКОГО"

КАФЕДРА МАТЕМАТИЧНИХ МЕТОДІВ СИСТЕМНОГО АНАЛІЗУ

КУРСОВА РОБОТА

з дисципліни

«Методи та технології обчислювального інтелекту»

на тему «Аналіз кредитних ризиків для фізичних осіб в умовах невизначеності з використанням нечітких нейронних мереж»

Виконав:

студент 5-го курсу

групи КА-02мп

Літвинчук А.М.

Прийняв:

професор, д.т.н.

Зайченко Ю. П.

Оцінка \_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_» \_\_\_\_\_\_\_\_ 2020 р.

Київ - 2021

**ЗМІСТ**

[ВСТУП 3](#_Toc71188200)

[ПОСТАНОВКА ЗАДАЧІ 3](#_Toc71188201)

[ТЕОРЕТИЧНА ЧАСТИНА 5](#_Toc71188202)

[Нечітка нейронна мережа NEFCLASS 5](#_Toc71188203)

[Архітектура 5](#_Toc71188204)

[Алгоритм навчання 9](#_Toc71188205)

[Створення бази правил 10](#_Toc71188206)

[Навчання нечітких множин 11](#_Toc71188207)

[Метрики 12](#_Toc71188208)

[Точність 12](#_Toc71188209)

[F1 міра 13](#_Toc71188210)

[Крос-валідація 14](#_Toc71188211)

[ОПИС ПРОГРАМНОГО ПРОДУКТУ 16](#_Toc71188212)

[ОПИС ВХІДНИХ ДАНИХ 19](#_Toc71188213)

[ВИСНОВКИ 24](#_Toc71188214)

[СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ 25](#_Toc71188215)

[ДОДАТОК А. ЛІСТИНГ 26](#_Toc71188216)

# ВСТУП

Автоматизація прийняття рішень у наші часи добралась до різних сфер. Це економить час експерту та мінімізує можливість людської помилки. У даній роботі ми розглянемо банкову сферу, а саме скорингові моделі.

Алгоритми кредитного скорингу, які дозволяють здогадуватися про ймовірність дефолту – це метод, який банки використовують для визначення того, чи слід надавати позику чи ні. У цій роботі задачею буде передбачення, чи зазнає клієнт фінансового банкрутства у найближчих 2 роки.

Існує безліч алгоритмів для прогнозування фінансового банкрутства, проте у даній роботі ми розглянемо саме нейронні мережі з нечіткою логікою. Їхньою перевагою є можливість враховувати невизначеність параметрів та обробка багатовимірних даних. Розглядатиметься саме модель NEFCLASS, яка об’єднує універсальні можливості нейронних мереж з можливістю обробки невизначеності даних. В результаті отримаємо модель, яка буде класифікувати фізичних осіб на 2 групи : на осіб, які будуть здатні оплатити кредит та осіб, які збанкрутують.

# ПОСТАНОВКА ЗАДАЧІ

Основними задачами цією роботи є:

* програмна реалізація нечіткої нейронної мережі NefClass та її навчання;
* програмна реалізації крос-валідаційного способу навчання та оцінки результатів моделі;
* імплементація метрик для оцінки результатів моделі;
* програмна реалізація перебору основних параметрів параметрів NefClass;
* оцінка результатів перебору параметрів, визначення найкращих варіантів, порівняльний аналіз та оцінка фінальної моделі.

# ТЕОРЕТИЧНА ЧАСТИНА

## Нечітка нейронна мережа NEFCLASS

Пропонується розглянути систему нечіткої класифікації NEuro-Fuzzy CLASSifier, скорочено надалі NEFCLASS, що заснована водночас на архітектурі нечіткого багатошарового персептрона та алгоритмах нечіткого логічного висновку. Зауважимо, що нечіткість даного алгоритму пояснюється недосконалістю або неповнотою вимірювань тих властивостей об'єктів, за якими здійснюється процес класифікації.

Модель NEFCLASS є нейро-нечітким класифікатором, принцип роботи якого заснований на отриманні нечітких правил з безлічі даних, які можна розділити на кілька непересічних класів.

Завдання NEFCLASS полягає в тому, щоб визначити приналежність до класу зразка з набором вхідних параметрів. Однак одна з головних проблем при проектуванні нейро-нечіткої системи полягає у правильні побудові функцій належності та визначенні нечітких правил.

Далі пропонується розглянути більш детально архітектуру нейро-нечіткої мережі NEFCLASS та покроково розібрати її алгоритм навчання.

### Архітектура

Нейро-нечітка мережа NEFClASS має тришарову послідовну архітектуру, до якої входить:

1. вхідний шар;
2. прихований або так званий шар нейронів нечітких правил;
3. вихідний шар нейронів класифікації.

Приклад схематичного зображення архітектури ННМ NEFClASS для задачі з двома входами, п’ятьма лінгвістичними правилами та двома класами подано на рис. 2.1., а кожен із зазначених вище шарів позначено як U1, U2 та U3 відповідно.

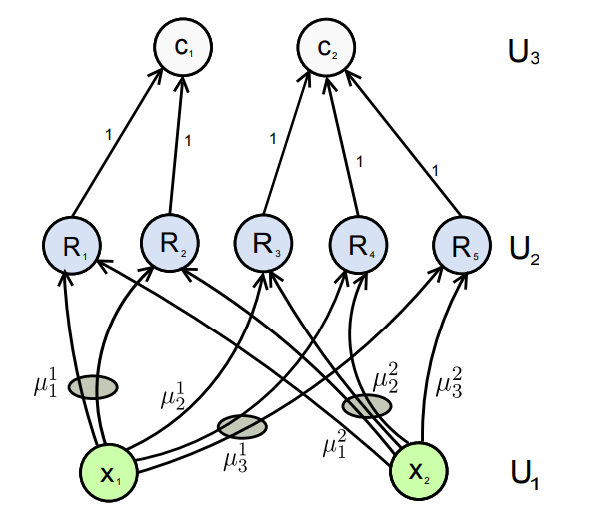
****

Рис. 2.1. – Архітектура нейро-нечіткої мережі NEFCLASS

Далі перейдемо до окремого розгляду кожного з трьох шарів мережі NEFClASS та зазначимо їх характерні особливості.

Тож, перший шар представляється вхідними нейронами, які складаються з вхідних зразків. Вхідні зразки, або як ще їх називають патерни – це вектори ознак певного об’єкта, що записуються як .Величина кожного з таких вхідних зразків представляється нечіткою множиною, при чому для кожної 𝑥𝑖 існує 𝑞𝑖 нечітких множин µ1𝑖 , … , µ𝑖𝑞𝑖.Тобто відбувається процес фазифікації вхідних нейронів, результатом якого є перехід від чітких значень патернів до нечітких значень відповідних лінгвістичних змінних. Таким чином для всіх вхідних змінних визначаються конкретні значення функції приналежності для кожної з лінгвістичних змінних. Їх можна записати у вигляді:

= , (1)

де jj – це той терм, який відповідає максимальному значенню функції належності i-вхідної змінної.

Кожна нечітка множина позначається лінгвістичним термом , що може, наприклад, набувати таких значень: «низький», «достатній», «відмінний» тощо. Тобто лінгвістичний терм ( *є*{1,…, }) є зв’язною ланкою між *xi* *є* *U*1 та *є U*2. Також варто зазначити, що з'єднання, які виходять із одного і того ж вузла вводу *х* та мають однакові мітки й однакову вагу в будь-який час, називаються зв'язними з'єднаннями, а їх вага називається зваженою вагою.

Другим шаром архітектури нейро-нечіткої мережі NEFClASS є прихований шар, який представлений базою правил, що складається з *k* нечітких лінгвістичних правил *R1, …, Rn* та є так званою апроксимацією невідомої функції : → , що описує класифікаційну задачу, де (𝑥) = (, … , ) така, що = 1, = 0 (𝑗 {1, … , 𝑚}) , а *x*  належить класу .

Активація нейронів цього шару з патерном *p* розраховується як функція *t-норми:*

= , (2)

де – це нечітка вага з’єднання вхідного нейрона з нейроном правила – це активація вхідного нейрона.

У загальному лінгвістичним правилом нечіткого логічного висновку *«якщо-то»* називається конструкція типу:

*R = якщо x є А, то у є В,* (3)

де *х* – це вхідна лінгвістична змінна,

*у* – це вихідна лінгвістична змінна,

*A* і *В* – це деякі нечіткі множини (функції приналежності), взяті із терм-множин змінних *x* та *y* відповідно.

Зазначимо, що у записі (3) вираз *«x є А»* – є нечітким висловлюванням, що характеризує умову правила (антецедент) а вираз *«у є В»* – позначає висновок правила (консеквент). Також, нечіткі множини, що направлені на однаковий вузол правил R, називаються антецедентами цього вузла.

Стосовно нейро-нечіткої мережі NEFCLASS, то нечіткі правила формуються наступним чином, реалізуючи висновок Мамдані:

(4)

Тобто це можна трактувати так що, *якщо і і … і , то патерн належить класу i.*

Тож, задача NEFCLASS полягає у тому, щоб визначити належність до класу *C ⊂*  вхідної ознаки .

Останній шар мережі складається з вихідних нейронів кожного класу, активація яких обчислюється на основі активації правил, які в якості висновку вказують на певний клас. Для цього використовується одна з формул, поданих нижче*:*

, або (5)

(6)

де *W(R, с)* – це нечітка вага з’єднання нейрона правил *R* з вихідним нейроном,

– це активація нейрона правил.

Як правило, *W(R, с)* дорівнює 1. Синаптичні ваги між нейроном правила та нейроном вихідного шару отримають значення 1, якщо поточний зразок навчальної вибірки дійсно належить цьому класу, і значення 0 для ваг з іншими класами.

### Алгоритм навчання

Результати роботи нейро-нечіткої мережі NEFCLASS напряму залежать від нечітких множин та лінгвістичних правил, що будуть застосовані до вхідних зразків. Вони обираються із множини вибірок шляхом навчання системи. Перед тим як розглянути сам процес навчання варто зазначити, що експерт має задати два обов’язкових параметри:

1. кількість початкових нечітких множин;
2. максимальна кількість вузлів правил, що можуть бути створені у прихованому шарі.

Зазначимо, що нейро-нечітка мережа NEFClass буде складатися з *n* вхідних нейронів (𝑥1, 𝑥2, ..., 𝑥𝑛), максимальної кількості нечітких правил (𝑘 ≤ 𝑘𝑚𝑎𝑥) та *m* вихідних нейронів (c1, c2, …, c𝑚). Окрім цього буде задано навчальну множину зразків *L* = {(𝑝1, 𝑡1), … , (𝑝𝑠, 𝑡𝑠)}, кожен з яких буде складатися з вхідного зразка та бажаного зразка *t ∈ .* Алгоритм навчання мережі NEFCLASS складається з двох етапів:

1. генерації або задання бази лінгвістичних правил;
2. налаштування ваг багатошарового персептрона за допомогою емпіричного алгоритму чи теоретично обґрунтованих алгоритмів навчання нейронних мереж.

Надалі пропонуємо розглянути кожен із цих етапів більш детально.

#### Створення бази правил

Метою даного етапу є створення *k-кількості* правил для мережі NEFCLASS. Задля цього потрібно:

1. обрати патерн (𝑝, 𝑡) із навчальної множини *L;*
2. для кожного вхідного нейрона *xi* *є U1* знайти таку функцію належності , що:

(7)

1. якщо *k-кількість* вузлів правил є меншою за максимально встановлену і не існує вузла правила *R* з умовою, що *W(x1, R)* = то створюємо такий вузол і з'єднуємо його з вихідним вузлом *c1,* якщо *t1 = 1;*
2. якщо ще лишилися неопрацьовані патерни із навчальної множини *L* та виконується умова, що 𝑘 ≤ 𝑘𝑚𝑎𝑥 , то переходимо до першого кроку. Інакше алгоритм перебору патернів зупиняється.

Наступним кроком є визначення безпосередньо самої бази правил, що може бути сформована за однією з трьох процедур:

1. *просте* навчання правил, тобто коли до бази правил потрапляють перші 𝑘 правил;
2. *найкраще* навчання правил, яке зводиться до наступного: при надходженні всіх зразків навчальної вибірки для кожного антецедента правила підраховується рейтинг:

де – це активація нейрона правил R.

Коефіцієнт може приймати два значення: «1» – у випадку, якщо зразок *p* класифіковано правильно та «-1» – за умови хибної класифікації. Потім з усіх нейронів правил мережі NEFCLASS обираємо і залишаємо в базі *k* нейронів правил з найвищими значеннями .

1. *найкраще для кожного класу* навчання правил подібний до *найкращого* навчання правил за винятком того, що для кожного класу залишаються ті кращі [ ] правил, наслідки яких представляють клас (де [𝑥] − ціла частина від 𝑥).

#### Навчання нечітких множин

Під час цього етапу проводитьсянавчання параметрів функцій належності нечітких множин. Алгоритм навчання з учителем повинен адаптувати нечіткі множини, пробігаючи циклічно через усю навчальну множину *L* до тих пір поки не спрацює один із критеріїв зупинки. Надалі sрозглянемо усі етапи цього алгоритму детальніше:

1. обираєм патерн (𝑝, 𝑡) із навчальної множини *L* і подаємо його на вхід мережі, визначаючи вихідний вектор *с;*
2. для кожного вихідного нейрона *ci* визначаємо значення похибки:

де – це бажаний вихід нейрона,

– це фактичний вихід нейрона *ci*.

1. для кожного нейрона правил R, для якого виконується умова, що вихід

> 0, виконуємо наступні розрахунки:

* обчислюємо значення похибки нейрона правил R:
* знаходимо такий вхідний нейрон 𝑥′ що

𝑊(𝑥′, 𝑅) (𝑎𝑥′ ) =

* для нечітких множин 𝑊(𝑥′, 𝑅) розраховуємо зміщення параметрів 𝛿𝑎, 𝛿𝑏 та 𝛿𝑐 функції приналежності, використовуючи параметр – швидкість навчання 𝜎 > 0:

І за одержаними розрахунками змінюємо 𝑊(𝑥′, 𝑅).

* обчислюємо похибку правила:

Алгоритм навчання параметрів функцій належності зупиняється якщо спрацював один із критеріїв зупинки, наприклад:

1. похибка набуває значення близького до нуля;
2. похибка не зменшується протягом *n* ітерацій.

## Метрики

### Точність

Основна метрика, яка часто використовується у задачах класифікації. Має таке визначення:

Accuracy = (TP + TN) / (TP + FP + TN + FN)

де TP – кількість вгаданих прикладів з класом 1;

TN - кількість вгаданих прикладів з класом 0;

FP – кількість помилок, коли модель дала клас 1, а насправді там 0;

FN – кількість помилок, коли модель дала клас 0, а нсправді там 1.

Проте ця метрика є дуже оманливою. Скажімо, у нас 99 прикладів класу 0 та 1 приклад класу 1. Уявімо, що наша модель будь-якому прикладу присвоєю мітку 0. Тоді у нас точність:

Accuracy = (0 + 99) / (0 + 0 + 99 + 1) = 99 / 100 = 99%

Здається, що точність аж 99%, проте ми не вгадали жодного прикладу з класом 1, що є дуже погано наприклад у випадку важких хвороб. Аналогічно і в нас, нам не потрібно щоб модель завжди казала, що людина збанкрутує або навпаки, що людина завжди віддаватиме борг. Тому введемо ще одну метрику.

### F1 міра

Введемо визначення цієї метрики:

Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

F1 = 2 \* Precision \* Recall / (Precision + Recall)

де TP, FP, FN – те саме, що у точності;

Precision – допоміжна метрика, яка показує відсоток вгаданих прикладів з тих, яким ми дали клас 1;

Recall – допоміжна метрика, яка показує відсоток вгаданих прикладів з усіх прикладів з класом 1 (повнота).

Розглянемо попередній іграшковий приклад:

Precision = 0 / (0 + 0 + 1e-9) = 0 (в даному випадку у знаменнику додаємо маленький епсілон, тому результат 0)

Recall = 0 / (0 + 1) = 0

F1 = 2 \* 0 \* 0 / (0 + 0 + 1e-9) = 0

Тобто цей варіант чесно оцінює нашу модель, крім того він повністю незалежний від незбалансованості класів. У експериментах будемо моніторити обидві метрики, проте орієнтуватись будемо саме на F1-міру.

## Крос-валідація

Кожен експеримент потрібно оцінити за певною метрикою. Є декілька варіантів як отримати значення метрики на тестовій вибірці, один з них – просто розбити дані, що ми маємо на тренувальну та тестову вибірки, навчати на тренувальній, а валідаувати відповідно на тестовій. Проте у цього способу є недостатки, якщо ми хочемо підбирати гіперпараметри мережі, а саме висока ймовірність перенавчання під цю тренувальну та тестові вибірки, тобто результати з кожним експериментом будуть все менш об’єктивиними та збільшується шанс, що підібрані параметри будуть погано переноситись і на інші приклади з даного розподілу даних.

Саме тому використовувати ми будемо механізм під назвою крос-валідація. У неї є параметри – кількість розбиттів. У даній роботі розглядається 5 розбиттів. Алгоритм крос-валідації наступний:

1. Випадково (або за певним принципом) беремо (N – 1) / N відсотків даних у тренувальну вибірку, 1 / N відсотків даних у тестову (при цьому тестові дані завжди мають буть унікальними для кожного розбиття, тобто перетину немає)
2. Навчаємо модель на тренувальній вибірці
3. Валідуємо модель на тестовій вибірці, зберігаємо метрики
4. Робимо пункти 1-3 N разів (тестова вибірка кожного разу має бути унікальна і без перетинів між розбиттями)
5. Усереднюємо отримані метрики

Цей процес можна зобразити так:

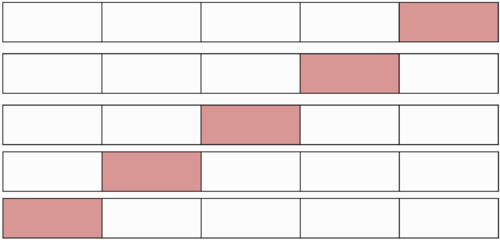


Рис. 1 Крос-валідація, білий блочок – дані, які йдуть у тренувальну вибірку, червоний – у тестову

Завдяки крос-валідації ми можемо отримати більш точну оцінку моделі, оскільки маємо тепер N спостережень (у нашому випадку 5). Всі моделі ми будемо навчати і оцінювати саме за таким принципом.

# ОПИС ПРОГРАМНОГО ПРОДУКТУ

Програмний продукт був розроблений на мові програмування Python. Весь код доступний за посиланням : <https://github.com/litvinich/mtoi_coursework>.

1. Перш за все був розроблений клас, який містив реалізацію мережі NEFCLASS та функцію для його навчання (знаходиться в src/neflcass.py та src/train\_utils.py)
2. Метрики були реалізовані за допомогою бібліотеки з відкритим кодом sklearn.
3. Після цього було розроблено функцію для проведення крос-валідації (знаходиться в src/train\_utils.py).
4. Було реалізовано швидкий перебір параметрів з легким інтерфейсом. Перебір параметрів відбувається паралельно на всіх ядрах процесора для максимального швидкого перебору параметрів (перебір параметрів у мене зайняв приблизно 4.5 години на 20 ядрах, без паралелізації це б зайняло приблизно 50-70 годин).

Розгляньмо інтерфейс основної функції для перебору параметрів (<https://github.com/litvinich/mtoi_coursework/blob/master/grid_search.py>):

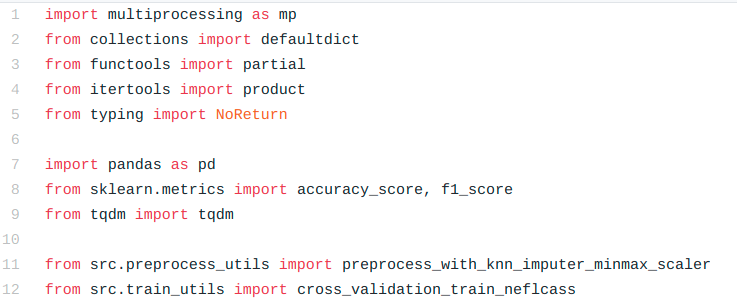


Рис.2 Підгрузка необхідних функцій так класів

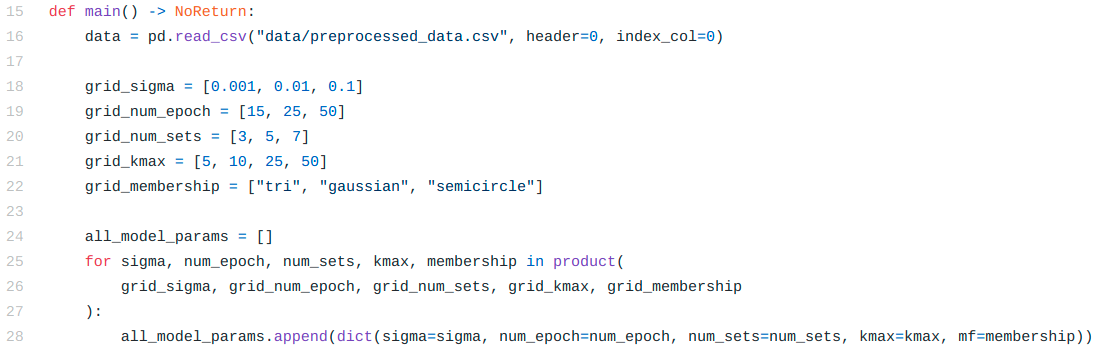


Рис.3 Підгрузка даних та визначення параметрів, які ми хочемо перебирати (можливо перебирати швидкість навчання, кількість епох, тобто довжина навчання, кількість початкових нечітких множин, максимальна кількість вузлів правил, функцію належності)

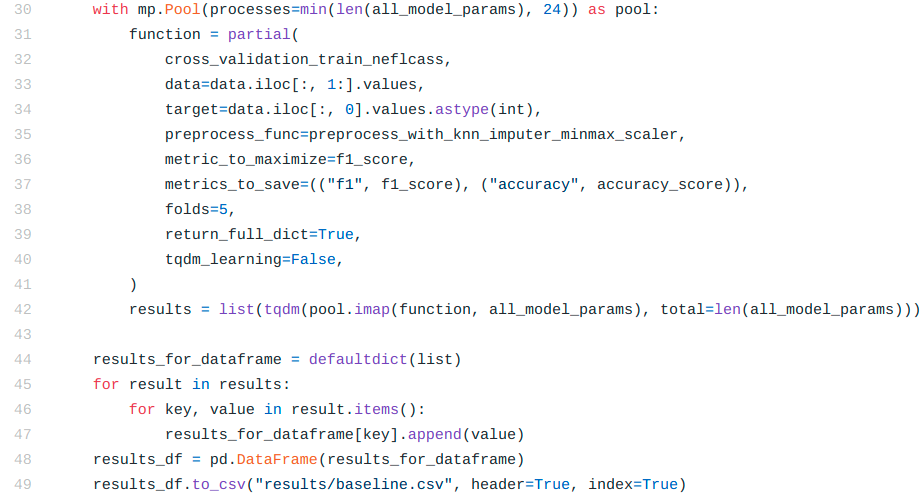


Рис.4 Паралельний запуск крос-валідації для кожного набору параметрів та збереження результатів у “results/baseline.csv”

Для запуску виконуємо наступну команду:

python grid\_search.py

В результаті отримаємо ось таку таблицю з результатами для подальшого аналізу:

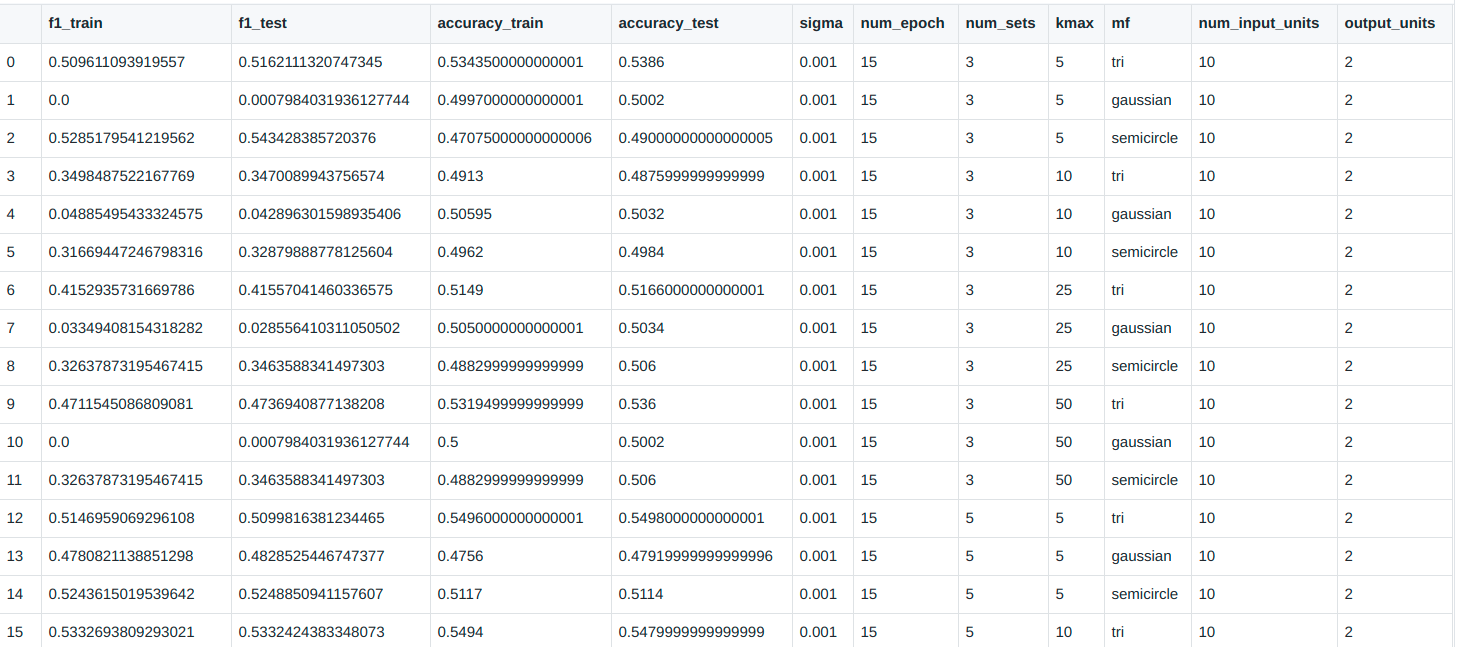


Рис. 5 Результати роботи

# ОПИС ВХІДНИХ ДАНИХ

Дані були взяті з сайту kaggle.com , а саме з змагання Give Me Some Credit (<https://www.kaggle.com/c/GiveMeSomeCredit>). Задача була досить важкою, оскільки для її вирішення залучили велику кількість бажаючих з грошовою винагородою за найкраще рішення. Проте початкова вибірка була занадто великою (більше 150000 кредиторів), тому я взяв підвибірку, опрацювання якої займало б розумний час на комп’ютері, що я маю – 5000 прикладів (2500 з них віддали борг за 2 роки, а 2500 – ні).

Задачею було реалізувати модель для передбачення чи збанкрутує фізична особа у найближчих 2 роки. Для цього пропонувалось використати такі предиктори:

1. RevolvingUtilizationOfUnsecuredLines – загальний залишок за кредитними картками та особистими кредитними лініями, за винятком нерухомості та відсутність боргу в розстрочку, як автокредити, поділений на суму кредитних лімітів (тип даних: відсоток).
2. Age – вік фізичної особли (тип даних: ціле число).
3. NumberOfTime30-59DaysPastDueNotWorse – кількість разів, коли позичальник прострочував 30-59 днів, але не більше за останні 2 роки (тип даних: ціле число).
4. DebtRatio – щомісячні виплати боргу, аліменти, витрати на життя, поділені на місячний валовий дохід (тип даних: відсоток).
5. MonthlyIncome – місячний дохід (тип даних: ціле число).
6. NumberOfOpenCreditLinesAndLoans – кількість відкритих позик (розстрочка, наприклад, автокредитування або іпотека) та кредитних ліній (наприклад, кредитних карток) (тип даних: ціле число).
7. NumberOfTimes90DaysLate – кількість випадків, коли позичальник прострочував 90 днів або більше (тип даних: ціле число).
8. NumberRealEstateLoansOrLines – кількість іпотечних кредитів та позик на нерухомість, включаючи кредитні лінії власного капіталу (тип даних: ціле число).
9. NumberOfTime60-89DaysPastDueNotWorse – кількість випадків, коли позичальник прострочував 60-89 днів, але не більше за останні 2 роки (тип даних: ціле число).
10. NumberOfDependents – кількість утриманців у сім'ї, крім них самих (дружина, діти тощо) (тип даних: ціле число).

Варто підмітити, що всі предиктори – цілі числа або відсотки, тобто ніякої додаткової обробки текстових змінних не потрібно, лише нормалізація. Проведемо більш детальний огляд розподілів цих величин, щоб зрозуміти складність задачі.

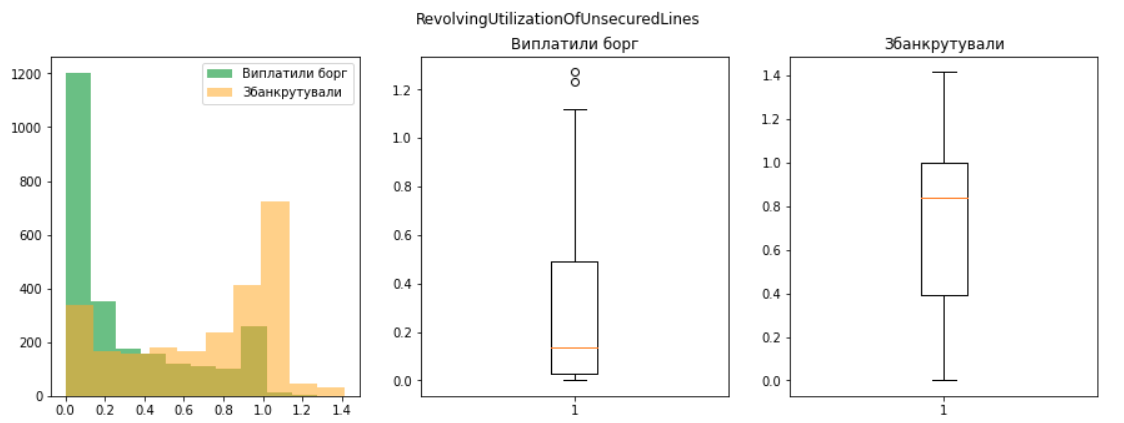


Рис. 6 Розподіли в залежності від класу предиктора 1

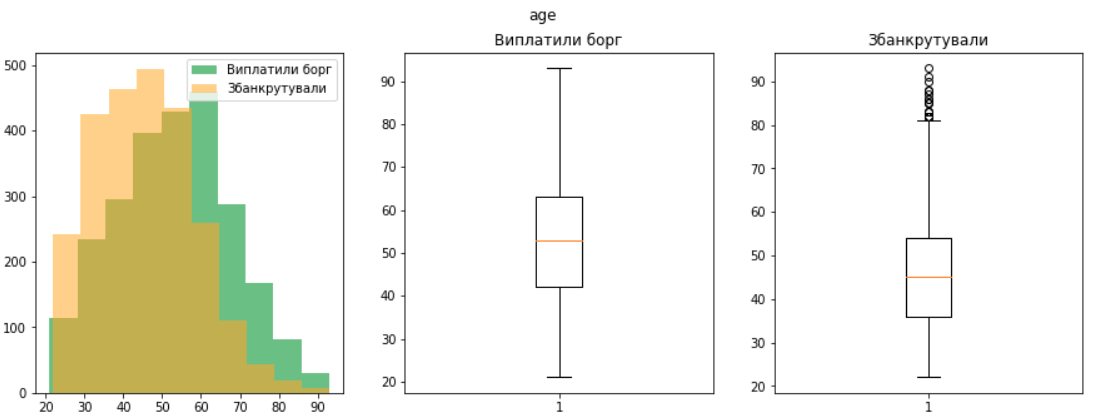


Рис. 7 Розподіли в залежності від класу предиктора 2

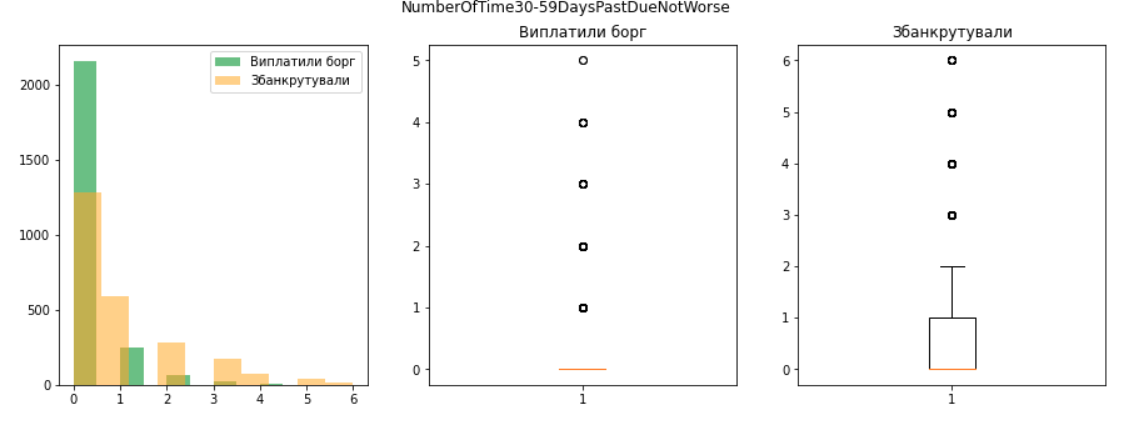


Рис. 8 Розподіли в залежності від класу предиктора 3

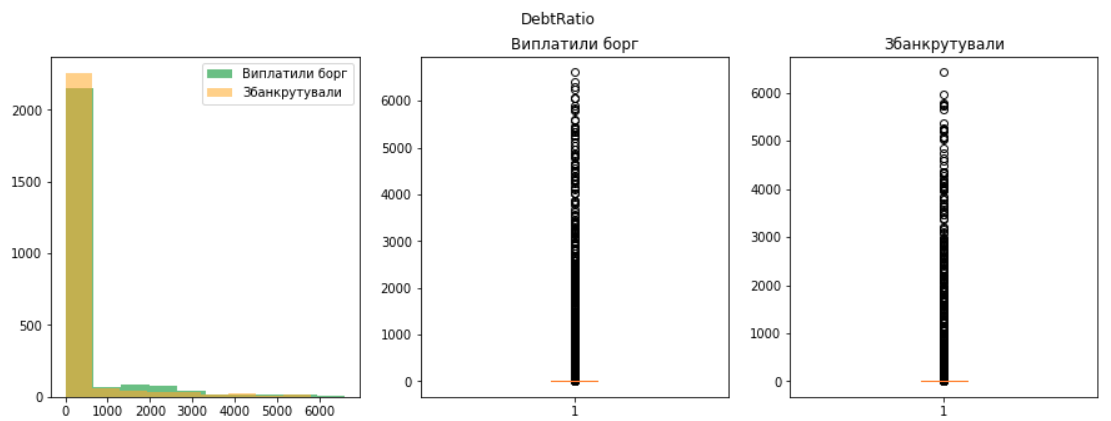


Рис. 9 Розподіли в залежності від класу предиктора 4

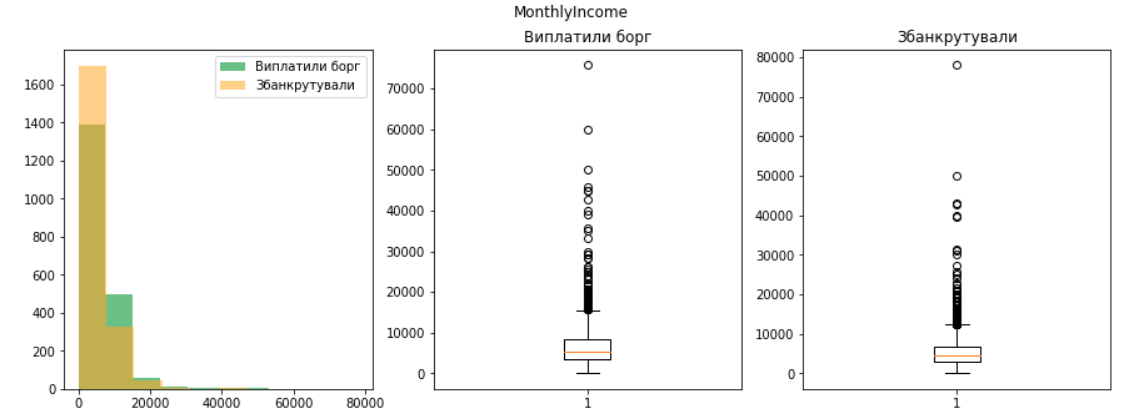


Рис. 10 Розподіли в залежності від класу предиктора 5

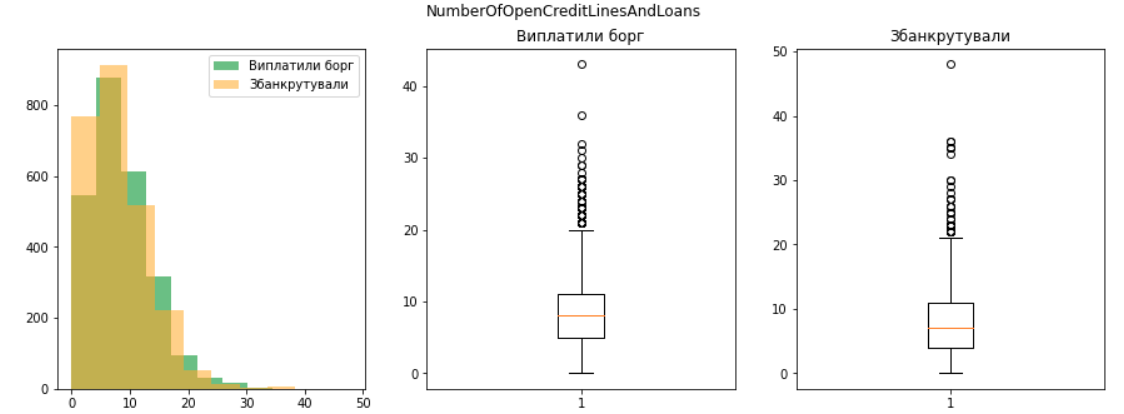


Рис. 11 Розподіли в залежності від класу предиктора 6

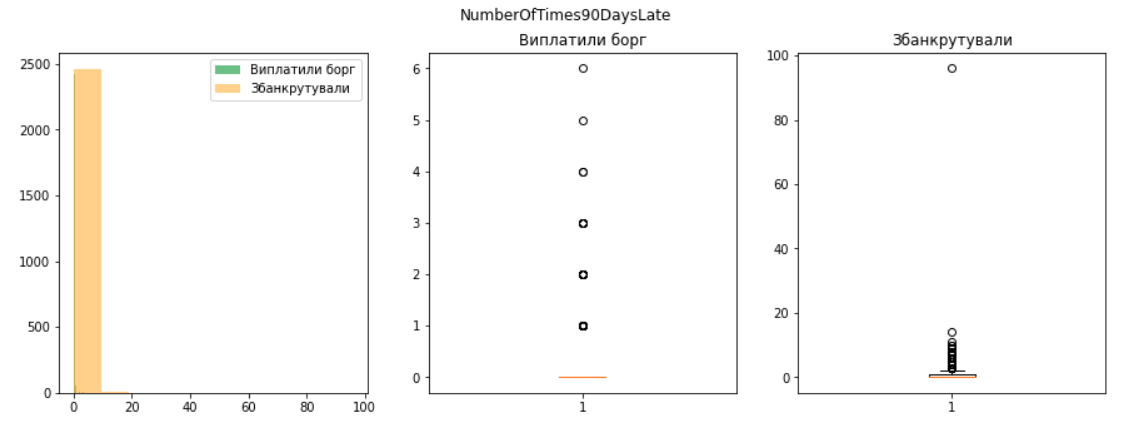


Рис. 12 Розподіли в залежності від класу предиктора 7

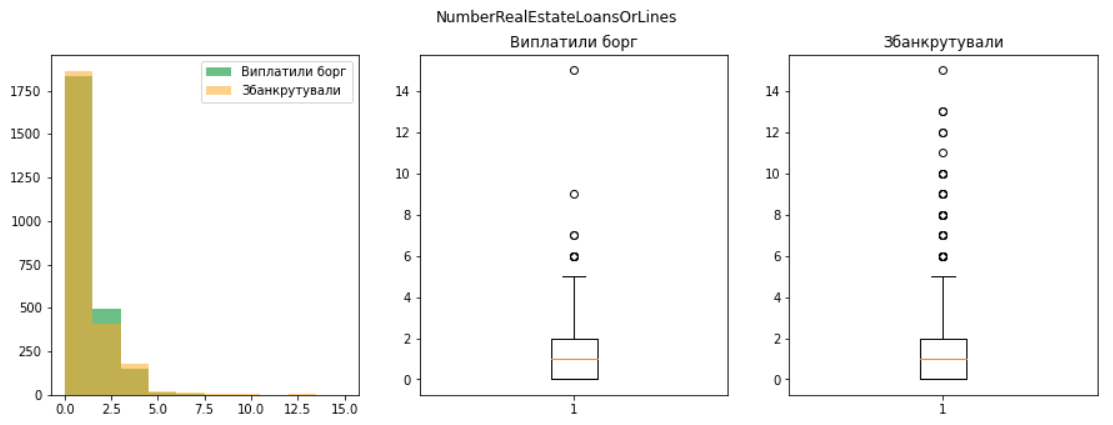


Рис. 13 Розподіли в залежності від класу предиктора 8

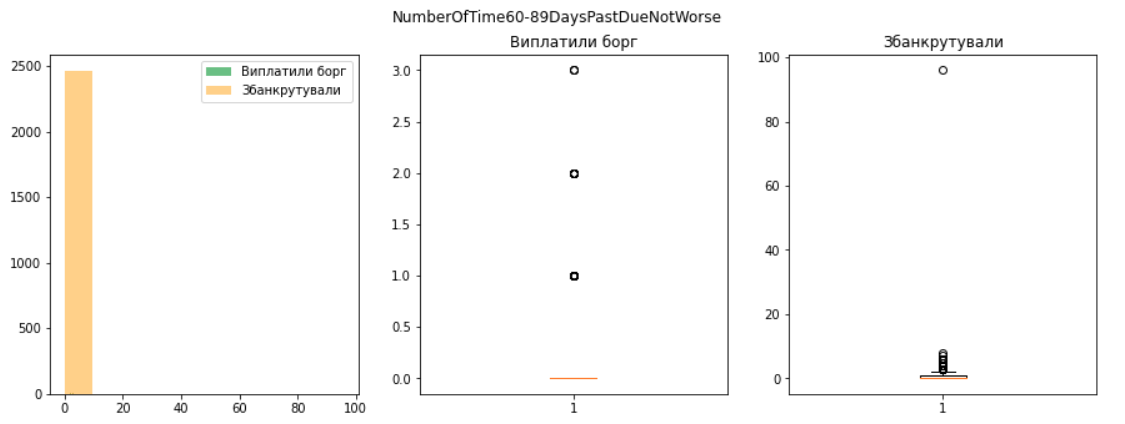


Рис. 14 Розподіли в залежності від класу предиктора 9

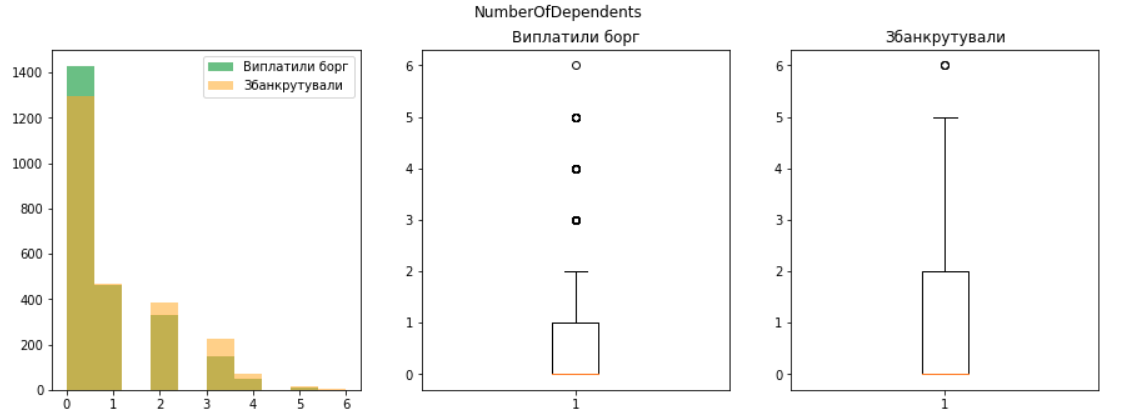


Рис. 15 Розподіли в залежності від класу предиктора 10

Деякі предиктори самі по собі несуть сильні ознаки майбутнього банкрутства (предиктори 1, 2, 3, 7, 9, 10), і це навіть між взаємних залежностей (людина з 3 дітьми і заробітком 3000 доларів набагато ймовірніше збанкрутує за людину з 3 дітьми та доходом 10000 доларів).

Таким чином ми побачили, що задача може вирішуватись з непоганою точністю (немає жодного предиктора, по якому можна було б ідеально відділити класи, тому ідеально задача не вирішиться, варто сподіватись на 60-70% точності).

# РЕЗУЛЬТАТИ РОБОТИ

В результаті було досліджено наступні параметри моделі:

grid\_sigma = [0.001, 0.01, 0.1]

grid\_num\_epoch = [15, 25, 50]

grid\_num\_sets = [3, 5, 7]

grid\_kmax = [5, 10, 25, 50]

grid\_membership = ["tri", "gaussian", "semicircle"]

Це відповідно швидкість навчання, кількість епох (довжина навчання), кількість початкових нечітких множин, максимальна кількість вузлів правил, функція належності.

Крос-валідація відбувалась на 5 розбиттях. Під час навчання зберігалась модель, яка є найкращою по цільовій метриці. Цільову метрику можна обрати – це може бути точність або F1-міра. Було досліджено обидва методи, щоб показати мінуси точності як цільової метрики. Якщо цільова метрика не покращувалась протягом 20 епох – зупиняли навчання (early stopping).

Спочатку оберемо найкращий варіант для кожного параметру. Щоб це зробити, зафіксуємо всі експерименти з цим параметром та візьмемо середнє для кожної метрики. Таким чином отримаємо справедливу оцінку і не буде варіанту, що ми обрали певний параметр через випадковий збіг обставин.

# ВИСНОВКИ

У даній курсовій роботі був розроблений прототип системи підтримки прийняття рішення для вибору оптимального фільтру, який відповідає поставленій задачі. Були визначені майбутні шляхи розвитку та покращення даного ПП.

# СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. <https://www.kaggle.com/timoboz/tesla-stock-data-from-2010-to-2020>
2. Бідюк П.І., Коршевнюк Л.О. Проектування комп'ютерних інформаційних систем підтримки прийняття рішень: навч. посіб. / ННК «Інститут прикладного системного аналізу» Національний технічний університет України «Київський політехнічний інститут», 2010. 340 с.
3. Синицын И.Н. Фильтры Калмана и Пугачева, 2006. 640 с.
4. Дрейпер Н., Смит Г. Прикладной регресионный анализ, 2007. 912 с.
5. Forsyth D. Applied Machine Learning, 2019. 478 p.

# ДОДАТОК А. ЛІСТИНГ

Весь код доступний на <https://github.com/litvinich/smpr_course_project>

Файл app.py

import sys

from typing import Any, NoReturn, Optional

import pandas as pd

from matplotlib import pyplot as plt

from PyQt5.QtGui import QFont

from PyQt5.QtWidgets import (

QApplication,

QCheckBox,

QComboBox,

QFileDialog,

QFrame,

QLabel,

QLineEdit,

QPushButton,

QTextEdit,

QVBoxLayout,

QWidget,

)

from models.model import BestFilterFinder

class App(QWidget):

def \_\_init\_\_(self) -> None:

super().\_\_init\_\_()

self.model\_name = "LinearRegression"

self.metric\_name = "MAE"

self.validation\_percent\_int: float = 5

self.init\_ui()

def init\_ui(self) -> NoReturn:

font\_bold = QFont()

font\_bold.setBold(True)

up = QFrame(self)

up.setFrameShape(QFrame.StyledPanel)

label\_input\_data = QLabel("Вхідні дані", up)

label\_input\_data.move(2, 10)

self.input\_data = QLineEdit(up)

self.input\_data.setText("./data/tesla\_stock.csv")

self.input\_data.setFixedWidth(200)

self.input\_data.move(100, 7)

open\_input\_data = QPushButton("...", up)

open\_input\_data.setCheckable(True)

open\_input\_data.move(310, 7)

open\_input\_data.clicked[bool].connect(self.open\_input\_data\_dialog)

label\_target\_column = QLabel("Цільова змінна", up)

label\_target\_column.move(2, 50)

self.input\_target\_column = QLineEdit(up)

self.input\_target\_column.setText("Close")

self.input\_target\_column.setFixedWidth(100)

self.input\_target\_column.move(145, 45)

label\_index\_column = QLabel("Колонка індексів", up)

label\_index\_column.move(2, 90)

self.index\_column = QLineEdit(up)

self.index\_column.setText("0")

self.index\_column.setFixedWidth(30)

self.index\_column.move(160, 86)

label\_p = QLabel("P", up)

label\_p.move(2, 135)

self.p = QLineEdit(up)

self.p.setText("10")

self.p.setFixedWidth(30)

self.p.move(20, 130)

label\_q = QLabel("Q", up)

label\_q.move(60, 135)

self.q = QLineEdit(up)

self.q.setFixedWidth(30)

self.q.move(80, 130)

validation\_percent\_label = QLabel("Відсоток даних для валідації", up)

validation\_percent\_label.move(2, 180)

self.validation\_percent = QLineEdit(up)

self.validation\_percent.setText("5")

self.validation\_percent.setFixedWidth(40)

self.validation\_percent.move(255, 175)

model\_label = QLabel("Модель", up)

model\_label.move(450, 10)

self.model\_name\_box = QComboBox(up)

self.model\_name\_box.addItems(

[

"LinearRegression",

"RidgeRegression",

"LassoRegression",

"SVM",

"XGBoostRegression",

"RandomForestRegression",

]

)

self.model\_name\_box.move(530, 7)

self.model\_name\_box.activated[str].connect(self.model\_name\_handler)

metric\_label = QLabel("Метрика для оптимізації", up)

metric\_label.move(450, 50)

self.metric\_name\_box = QComboBox(up)

self.metric\_name\_box.addItems(["MAE", "MSE", "R2"])

self.metric\_name\_box.move(670, 47)

self.metric\_name\_box.activated[str].connect(self.metric\_name\_handler)

self.plot\_checkbox = QCheckBox("Малювати графік", up)

self.plot\_checkbox.setChecked(True)

self.plot\_checkbox.move(450, 85)

self.execute\_button = QPushButton("Виконати", up)

self.execute\_button.move(550, 230)

self.execute\_button.clicked.connect(self.execute)

table\_frame = QFrame(self)

table\_frame.setFrameShape(QFrame.StyledPanel)

self.output = QTextEdit(table\_frame)

self.output.setReadOnly(True)

self.output.setLineWrapMode(QTextEdit.NoWrap)

self.output.setFixedWidth(1160)

self.output.setMinimumHeight(300)

self.output.setMaximumHeight(1000)

self.output.move(5, 5)

vertical\_layout = QVBoxLayout()

vertical\_layout.addWidget(up)

vertical\_layout.addWidget(table\_frame)

self.setLayout(vertical\_layout)

self.setGeometry(150, 150, 1200, 600)

self.setWindowTitle("Порівняння методів фільтрації")

self.show()

def open\_input\_data\_dialog(self) -> NoReturn:

options = QFileDialog.Options()

options |= QFileDialog.DontUseNativeDialog

target\_path = str(QFileDialog.getOpenFileName(self, "Виберіть файл", options=options)[0])

if target\_path:

self.input\_data.setText(target\_path)

@staticmethod

def text\_to\_int(text: str, default: Optional[Any] = None) -> Optional[Any]:

try:

return int(text)

except ValueError:

return default

def model\_name\_handler(self, value: str) -> NoReturn:

self.model\_name = value

def metric\_name\_handler(self, value: str) -> NoReturn:

self.metric\_name = value

@staticmethod

def percent\_handler(value: str) -> float:

int\_value = int(value)

if 0 < int\_value <= 1:

return int\_value

elif 0 < int\_value <= 100:

return int\_value \* 0.01

else:

raise ValueError("Такий відсоток не підтримується")

def execute(self) -> None:

try:

data: pd.DataFrame = pd.read\_csv(

self.input\_data.text(), index\_col=App.text\_to\_int(self.index\_column.text())

)

except FileNotFoundError:

print("Неправильний шлях до файлу")

return

except IndexError:

print("Такої колонки індексів не існує")

return

try:

variable: pd.Series = data[self.input\_target\_column.text()]

except KeyError:

print("Такої цільової змінної не існує")

return

try:

self.validation\_percent\_int = App.percent\_handler(self.validation\_percent.text())

except ValueError:

print("Такий відсоток не підтримується")

model = BestFilterFinder(

model\_name=self.model\_name,

metric\_name=self.metric\_name.lower(),

validation\_percent=self.validation\_percent\_int,

)

text = ""

y\_test, ma\_filter, ma\_predict, ma\_params, ma\_metrics = model.grid\_search\_moving\_average(

variable=variable.copy(), q=App.text\_to\_int(self.q.text()), p=App.text\_to\_int(self.p.text(), default=1)

)

\_, exp\_ma\_filter, exp\_ma\_predict, exp\_ma\_params, exp\_ma\_metrics = model.grid\_search\_exp\_moving\_average(

variable=variable.copy(), q=App.text\_to\_int(self.q.text()), p=App.text\_to\_int(self.p.text(), default=1)

)

\_, kalman\_filter, kalman\_predict, kalman\_params, kalman\_metrics = model.grid\_search\_kalman(

variable=variable.copy(), q=App.text\_to\_int(self.q.text()), p=App.text\_to\_int(self.p.text(), default=1)

)

text += (

f"Moving Average: {ma\_metrics} з параметрами q = {ma\_params.q} та вікном = {ma\_params.moving\_average}\n"

)

text += f"Exponential Moving Average: {exp\_ma\_metrics} з параметрами q = {exp\_ma\_params.q} та alpha = {exp\_ma\_params.alpha}\n"

text += f"Kalman Filter: {kalman\_metrics} з параметрами q = {exp\_ma\_params.q}\n"

self.output.setText(text)

try:

y\_test.index = pd.to\_datetime(y\_test.index)

except ValueError:

pass

if self.plot\_checkbox.isChecked():

fig, axs = plt.subplots(2, 1, figsize=(16, 16))

axs[0].plot(y\_test.index, y\_test, label="Правильні значення")

axs[0].plot(y\_test.index, ma\_filter, label=f"Moving Average з вікном = {ma\_params.moving\_average}")

axs[0].plot(

y\_test.index, exp\_ma\_filter, label=f"Exponential Moving Average з alpha = {exp\_ma\_params.alpha}"

)

axs[0].plot(y\_test.index, kalman\_filter, label=f"Kalman Filter")

axs[0].set\_title("Візуалізація найкращих фільтрів")

axs[0].legend()

axs[1].plot(y\_test.index, y\_test, label="Правильні значення")

axs[1].plot(

y\_test.index,

ma\_predict,

label=f"{self.model\_name} та Moving Average з вікном = {ma\_params.moving\_average}",

)

axs[1].plot(

y\_test.index,

exp\_ma\_predict,

label=f"{self.model\_name} та Exponential Moving Average з alpha = {exp\_ma\_params.alpha}",

)

axs[1].plot(y\_test.index, kalman\_filter, label=f"{self.model\_name} та Kalman Filter")

axs[1].set\_title("Передбачення моделей, які використовують найкращі фільтри")

axs[1].legend()

plt.show()

def main() -> NoReturn:

pyqt\_application = QApplication(sys.argv)

\_ = App()

sys.exit(pyqt\_application.exec\_())

if \_\_name\_\_ == "\_\_main\_\_":

main()

Файл models/dataset.py

from typing import Optional

import numpy as np

import pandas as pd

def create\_ar\_filter\_table(

variable: pd.Series, p: int = 1, q: int = 0, filter\_variable: Optional[pd.Series] = None, filter\_name: str = "ma"

) -> pd.DataFrame:

columns = ["free\_coef"]

index = variable.index

values = np.zeros((len(variable), 1 + p + q))

# free\_coef

values[:, 0] = 1

# autoregression

for ar in range(1, p + 1):

columns.append(f"ar({ar})")

values[: (ar - 1), ar] = np.nan

values[(ar - 1) :, ar] = variable[: (len(variable) - ar + 1)]

if q > 0 and filter\_variable is not None:

for ma in range(1, q + 1):

columns.append(f"{filter\_name}({ma})")

values[: (ma - 1), p + ma] = np.nan

values[(ma - 1) :, p + ma] = filter\_variable[: (len(filter\_variable) - ma + 1)]

return pd.DataFrame(values, columns=columns, index=index)

def create\_next\_day\_price(variable: pd.Series) -> pd.Series:

variable[:-1] = variable[1:]

variable[-1] = np.nan

return variable

Файл models/model.py

import multiprocessing as mp

from collections import namedtuple

from functools import partial

from itertools import product

from typing import Callable, Iterable, List, Optional, Tuple, Union

import numpy as np

import pandas as pd

from simdkalman import KalmanFilter

from sklearn.base import BaseEstimator

from sklearn.ensemble import RandomForestRegressor

from sklearn.linear\_model import Lasso, LinearRegression, Ridge

from sklearn.metrics import mean\_absolute\_error, mean\_squared\_error, r2\_score

from sklearn.svm import SVR

from tqdm import tqdm

from xgboost import XGBRegressor

from .dataset import create\_ar\_filter\_table, create\_next\_day\_price

MODEL\_MAPPING = {

"LinearRegression": LinearRegression(n\_jobs=-1),

"RidgeRegression": Ridge(),

"LassoRegression": Lasso(),

"SVM": SVR(),

"XGBoostRegression": XGBRegressor(n\_jobs=-1),

"RandomForestRegression": RandomForestRegressor(n\_jobs=-1),

}

MovingAverageGridParams = namedtuple("MovingAverageGridParams", ["q", "moving\_average", "p"])

ExpMovingAverageGridParams = namedtuple("ExpMovingAverageGridParams", ["q", "alpha", "p"])

KalmanGridParams = namedtuple("KalmanGridParams", ["q", "p"])

FinalMetrics = namedtuple("FinalMetrics", ["mae", "mse", "r2"])

GridParams = Union[MovingAverageGridParams, ExpMovingAverageGridParams, KalmanGridParams]

GridSearchResult = Tuple[pd.Series, pd.Series, np.ndarray, GridParams, FinalMetrics]

class BestFilterFinder:

def \_\_init\_\_(self, model\_name: str, metric\_name: str, validation\_percent: float, processes: int = 10):

self.\_validation\_percent = validation\_percent

self.\_model\_name = model\_name

self.\_metric\_name = metric\_name

self.\_metric\_maximize = self.\_metric\_name == "r2"

self.\_processes = processes

def \_load\_model(self) -> BaseEstimator:

return MODEL\_MAPPING.get(self.\_model\_name, LinearRegression())

@staticmethod

def get\_scores(y\_true: Iterable, y\_predict: Iterable) -> FinalMetrics:

mae = mean\_absolute\_error(y\_true, y\_predict)

mse = mean\_squared\_error(y\_true, y\_predict)

r2 = r2\_score(y\_true, y\_predict)

return FinalMetrics(mae=mae, mse=mse, r2=r2)

@staticmethod

def get\_moving\_average\_filter(variable: pd.Series, grid\_params: MovingAverageGridParams) -> pd.Series:

return variable.rolling(window=grid\_params.moving\_average).mean()

@staticmethod

def get\_exp\_moving\_average\_filter(variable: pd.Series, grid\_params: ExpMovingAverageGridParams) -> pd.Series:

return variable.ewm(alpha=grid\_params.alpha).mean()

@staticmethod

def get\_kalman\_filter(variable: pd.Series, grid\_params: KalmanGridParams) -> pd.Series:

kalman = KalmanFilter(

state\_transition=np.array([[1, 1], [0, 1]]),

process\_noise=np.diag([0.1, 0.01]),

observation\_model=np.array([[1, 0]]),

observation\_noise=1.0,

)

smoothed = kalman.smooth(variable)

return pd.Series(smoothed.states.mean[:, 0], index=variable.index)

def \_train\_and\_evaluate(

self, grid\_params: GridParams, variable: pd.Series, get\_filter\_method: Callable

) -> Tuple[np.ndarray, FinalMetrics]:

filter\_variable = get\_filter\_method(variable=variable, grid\_params=grid\_params)

x = create\_ar\_filter\_table(

variable=variable, p=grid\_params.p, q=grid\_params.q, filter\_variable=filter\_variable

)

x["next\_day\_price"] = create\_next\_day\_price(variable=variable)

x\_train, x\_test = (

x.iloc[: int(len(variable) \* (1 - self.\_validation\_percent))],

x.iloc[int(len(variable) \* (1 - self.\_validation\_percent)) :],

)

x\_train = x\_train.dropna()

x\_test = x\_test.dropna()

x\_train, y\_train = x\_train.iloc[:, :-1], x\_train.iloc[:, -1]

x\_test, y\_test = x\_test.iloc[:, :-1], x\_test.iloc[:, -1]

model = self.\_load\_model()

model.fit(x\_train, y\_train)

y\_predict = model.predict(x\_test)

metrics = BestFilterFinder.get\_scores(y\_test, y\_predict)

return y\_predict, metrics

def \_grid\_search(

self, all\_variants: List[GridParams], variable: pd.Series, get\_filter\_method: Callable

) -> GridSearchResult:

with mp.Pool(processes=self.\_processes) as pool:

results = list(

tqdm(

pool.imap(

partial(self.\_train\_and\_evaluate, variable=variable, get\_filter\_method=get\_filter\_method,),

all\_variants,

),

total=len(all\_variants),

)

)

sorted\_results = sorted(

zip(all\_variants, results),

key=lambda x: getattr(x[1][1], self.\_metric\_name),

reverse=self.\_metric\_maximize,

)

best\_result = sorted\_results[0]

y\_test = create\_next\_day\_price(variable=variable).dropna()[

int(len(variable) \* (1 - self.\_validation\_percent)) :

]

best\_filter = get\_filter\_method(variable=variable, grid\_params=best\_result[0]).loc[y\_test.index]

best\_params, (best\_predict, best\_metrics) = best\_result

return y\_test, best\_filter, best\_predict, best\_params, best\_metrics

def grid\_search\_moving\_average(self, variable: pd.Series, p: int, q: Optional[int]) -> GridSearchResult:

q\_range = range(1, min(int(len(variable) \* 0.1), 105), 5) if q is None else [q]

moving\_average\_range = range(1, min(int(len(variable) \* 0.1), 105), 5)

all\_variants: List[GridParams] = [MovingAverageGridParams(q=0, moving\_average=0, p=p)]

for q, moving\_average in product(q\_range, moving\_average\_range):

all\_variants.append(MovingAverageGridParams(q=q, moving\_average=moving\_average, p=p))

return self.\_grid\_search(

all\_variants=all\_variants, variable=variable, get\_filter\_method=BestFilterFinder.get\_moving\_average\_filter

)

def grid\_search\_exp\_moving\_average(self, variable: pd.Series, p: int, q: Optional[int]) -> GridSearchResult:

q\_range = range(1, min(int(len(variable) \* 0.1), 105), 5) if q is None else [q]

alpha\_range = np.arange(0.01, 1, 0.04)

all\_variants: List[GridParams] = []

for q, alpha in product(q\_range, alpha\_range):

all\_variants.append(ExpMovingAverageGridParams(q=q, alpha=alpha, p=p))

return self.\_grid\_search(

all\_variants=all\_variants,

variable=variable,

get\_filter\_method=BestFilterFinder.get\_exp\_moving\_average\_filter,

)

def grid\_search\_kalman(self, variable: pd.Series, p: int, q: Optional[int]) -> GridSearchResult:

q\_range = range(1, min(int(len(variable) \* 0.1), 105), 5) if q is None else [q]

all\_variants: List[GridParams] = []

for q in q\_range:

all\_variants.append(KalmanGridParams(q=q, p=p))

return self.\_grid\_search(

all\_variants=all\_variants, variable=variable, get\_filter\_method=BestFilterFinder.get\_kalman\_filter

)