НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ
«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ
імені ІГОРЯ СІКОРСЬКОГО»

Фізико-математичний факультет
Кафедра математичного аналізу та теорії ймовірностей

«На правах рукопису»
УДК 519.2

До захисту допущено:
Завідувач кафедри
_____ Олег КЛЕСОВ
«___»_____2024 р.

Магістерська дисертація

на здобуття ступеня магістра

за освітньо-науковою програмою «Страхова та фінансова математика»

зі спеціальності 111 «Математика»

на тему: «Методи заповнення пропущених значень в масивах даних»

Виконала:
студентка VI курсу, групи ОМ-21мн
Оласюк Світлана Олексіївна _____

Науковий керівник:
доктор фізико-математичних наук,
доцент Розора Ірина Василівна _____

Рецензент:
кандидат фізико-математичних наук,
доцент Забула Дмитро Васильович _____

Засвідчую, що у цій магістерській дисертації немає запозичень з праць інших авторів без відповідних посилань.
Студентка _____

Київ – 2024 року

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

Фізико-математичний факультет
Кафедра математичного аналізу та теорії ймовірностей

Рівень вищої освіти – другий (магістерський)

Спеціальність – 111 «Математика»

Освітньо-наукова програма «Страхова та фінансова математика»

ЗАТВЕРДЖУЮ
Завідувач кафедри
_____ Олег Клесов
«___»_____2024 р.

ЗАВДАННЯ
на магістерську дисертацію студенту

Оласюк Світлані Олексіївні

1. Тема дисертації «Методи заповнення пропущених значень в масивах даних», науковий керівник дисертації Розора Ірина Василівна, доцент, доктор фізико-математичних наук, затверджені наказом по університету від «01» квітня 2024 р. № 1523-с.

2. Термін подання студентом дисертації: 11 червня 2024 р.

3. Об'єкт вивчення: масиви даних з пропущеними значеннями.

4. Предмет дослідження: види пропущених даних, методи імпутації пропущених значень.

5. Перелік завдань, які потрібно розробити:

1. Ознайомитись з літературою та основними поняттями.

2. Дослідити стандартні методи роботи з пропущеними даними: методом заміни середнім/медіаною, методом швидкої заміни, методом k найближчих сусідів.

3. Дослідити ансамблеві методи машинного навчання для класифікації та регресії: алгоритм випадкового лісу та алгоритм максимального градієнтного підсилення.

4. Дослідити застосування алгоритмів випадкового лісу та максимального градієнтного підсилення для імпутації.

5. Реалізувати обрані методи в середовищі RStudio.

6. На підставі отриманих результатів зробити висновки про доцільність застосування кожного з методів до імпутації даних і масиві.

7. Оформлення магістерської дисертації.

6. Орієнтовний перелік графічного (ілюстративного) матеріалу: 22 слайди.

7. Орієнтовний перелік публікацій:

1. Тези доповіді на XII Всеукраїнській науковій конференції молодих математиків, 9-11 травня 2024 року.
2. Тези доповіді на XIII Міжнародній науково-практичній конференції, 31 травня – 2 червня 2024 року.

8. Дата видачі завдання: 5 лютого 2024 року.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Ознайомитись з літературою та основними поняттями.	06.02.2024-01.03.2024	Виконано
2	Дослідити стандартні методи роботи з пропущеними даними: методом заміни середнім/медіаною, методом швидкої заміни, методом k найближчих сусідів.	02.03.2024-15.04.2024	Виконано
3	Дослідити ансамблеві методи машинного навчання для класифікації та регресії: алгоритм	16.04.2024-25.04.2024	Виконано

	випадкового лісу та алгоритм максимального градієнтного підсилення.		
4	Дослідити застосування алгоритмів випадкового лісу та максимального градієнтного підсилення для імпутації.	26.04.2024-05.05.2024	Виконано
5	Реалізувати обрані методи в середовищі RStudio.	06.05.2024-30.05.2024	Виконано
6	На підставі отриманих результатів зробити висновки про доцільність застосування кожного з методів до імпутації даних і масиві.	31.05.2024-01.06.2024	Виконано
7	Оформлення магістерської дисертації.	01.06.2024-07.06.2024	Виконано

Студент Світлана Оласюк

Науковий керівник Ірина Розора

РЕФЕРАТ

Магістерська дисертація містить 37 сторінок, 15 першоджерел та 22 слайди презентації. Структурно робота складається зі списку термінів, вступу, теоретичної частини, основної частини, висновків та переліку використаної літератури.

Відсутні значення є поширеною проблемою статистичних досліджень, дуже багато методів імпутації та їхніх модифікацій було розроблено для використання в медичній статистиці (як метод імпутації за допомогою алгоритму випадкового лісу) чи соціологічних опитуваннях (як метод швидкої заміни). Просте пропущені значення зустрічаються в найбільш різноманітних сферах, часто для адекватної оцінки ризиків збитків від природніх чи техногенних катастроф бракує інформації про суми, в яку оцінюється завдана шкода і кількість постраждалий, в нашому випадку для дослідження було обрано дані щодо техногенних і природних катастроф за останні 124 роки, статистику взято з сайту Центру досліджень епідеміології катастроф (CRED).

Метою роботи є дослідження методів заповнення пропусків в масивах даних та аналіз отриманих результатів для визначення переваг та недоліків кожного з методів та доцільність використання для обраного типу даних.

Ключові слова: імпутація, заміна пропущених даних, метод заміни середнім, метод заміни медіаною, метод швидкої заміни, метод k найближчих сусідів, метод випадкового лісу, метод максимального градієнтного підсилення, hot deck, kNN, Random Forest, MissForest, XGBoost, MCAR, MAR, MNAR.

ABSTRAST

The master's thesis: 37 pages, 15 primary sources and 22 presentation slides. The work consists of a list of terms, introduction, theoretical part, main part, conclusions and a list of primary sources.

Missing values are a common problem in statistical research, and many imputation methods and their modifications have been developed for use in medical statistics (for example, random forest imputation) or sociological surveys (for example, hot deck). However, missing values happen almost everywhere. Often there is a lack of information about the amount of damage and the number of victims to assess the risk of damage from natural or man-made disasters. For the study, we chose data on disasters for the last 124 years; these statistics are taken from the website of the Center for Research on the Epidemiology of Disasters (CRED).

The purpose of the paper is exploring the methods of filling missing values in datasets and analyzing the results to determine the advantages and disadvantages of each method and the feasibility of using it for our type of data.

Keywords: imputation, missing data, mean imputation, median imputation, hot deck imputation, k nearest neighbors imputation, random forest, eXtreme gradient boosting imputation, kNN, MissForest, XGBoost, MCAR, MAR, MNAR.

ЗМІСТ

СПИСОК ТЕРМІНІВ

Імпутація (англ. *imputation*) – це процес заміни відсутніх даних заміненими значеннями.

Hot deck – метод швидкої заміни.

kNN – метод k найближчих сусідів.

Random Forest – ансамблевий метод машинного навчання для класифікації та регресії даних, базується на алгоритмі беггінгу.

MissForest – метод випадкового лісу, що використовує алгоритм Random Forest для імпутації даних. Часто метод імпутації даних теж називають Random Forest.

XGBoost – ансамблевий метод машинного навчання для класифікації та регресії даних, базується на алгоритмі бустингу. Часто цим терміном також називають метод імпутації даних, що використовує XGBoost алгоритм.

ВСТУП

Науковцям часто доводиться мати справу з неповними даними, але, оскільки, більшість програм і аналітичних методів не передбачають наявності пропусків, виникла природня необхідність в розробці методик для усунення проблеми.

Пропущені дані були проблемою в наукових і соціологічних дослідженнях протягом століть, проте лише наприкінці 1980-х років було опубліковано дві основоположні книги, які заклали основу для прогресу в обробці відсутніх даних: «Статистичні дослідження з пропущеними значеннями» Родеріка Літтла і Дональда Рубіна (1987) та «Множинна імпутація пропущених даних при опитуваннях» (1987) Дональда Рубіна.

Доктор Родерік Літтл, професор біостатистики в Мічиганському університеті

Дональд Рубін, професор статистики в Гарвардському університеті

Пропущені значення – це дані, які невідомі для змінної чи змінних в деякому спостереженні. Проблема відсутніх даних актуальна маже для будь-якого виду дослідження або опитування і може мати значний вплив на висновки, які ми робимо і спричиняє багато проблем:

1. Відсутність даних зменшує статистичну якість, яка стосується ймовірності того, що тест відхилить нульову гіпотезу, якщо вона хибна.
2. Втрачені дані можуть спричинити зміщення в оцінці параметрів
3. Неповні дані погіршують репрезентативність вибірок.
4. Ускладнення аналізу результатів.

Відсутність частини інформації може бути як випадковою як і не випадковою. Наприклад, випадкові пропуски в анкетах можуть виникати як наслідок неуважності, недбалості частини опитуваних. Невипадкові пропущені значення можуть свідчити про небажання респондентів давати відповіді на чутливі питання.

В даній роботі ми не будемо заглиблюватись в причини виникнення проблем з повнотою даних та способів встановлення типу пропущених значень, нашою метою є розгляд видів пропущених даних та пошук практичних способів вирішення проблеми.

1 ВИДИ ПРОПУЩЕНИХ ЗНАЧЕНЬ

Перш ніж обрати метод заповнення пропущених в масиві значень важливо встановити чому вони відсутні. Згідно з класифікацією Д. Літтла і Р. Рубіна [1] відсутні значення можна поділити на три групи: пропущені повністю випадково (англ. *missing completely at random, MCAR*), пропущені не повністю випадково (англ. *missing at random, MAR*) та пропущені не випадково (англ. *not missing at random, MNAR*). Повністю випадкові пропуски, MCAR, рівномірно розподілені в масиві даних для всіх показників (на відміну від MAR).

Визначення виду пропусків є нетривіальною задачею, оскільки для однозначної відповіді потрібні були б, власне, відсутні дані. Можна встановити, що дані відсутні не випадково, але відрізнити MCAR від MAR дуже складно, якщо невідома причина неповноти інформації, наприклад, певна категорія опитуваних воліє пропускати незручні запитання в анкеті. У випадку невеликого відсотку пропусків можна вважати, що ми маємо справу з MCAR.

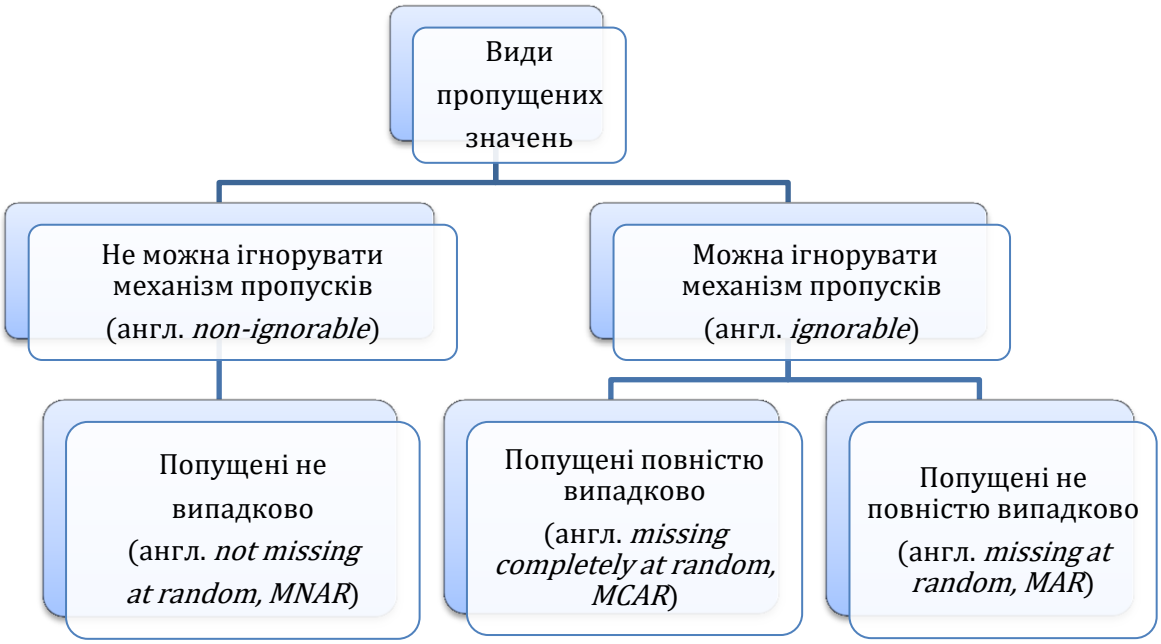

Рис. 1: Види пропущених значень

MNAR означає, що пропуски в масиві спричинені факторами, не пов'язаними з дослідженням, тобто не можна ігнорувати механізм пропусків. Відновлення даних, що пропущені зовсім невипадково в даній роботі не розглядається. Для вирішення проблеми пропусків, що класифікуються як MCAR або MAR існують різноманітні методи.

2 ТЕОРЕТИЧНА ЧАСТИНА

Перш за все варто зауважити, що найбільш очевидним рішенням є викидання з бази даних спостережень з неповними даними або ігнорування цих спостережень при аналізі чи розрахунках. Цей підхід може вимагати вилучення значної частини інформації, тому є не завжди можливим.

Всі методи імпутації поділяються на дві групи: методи одиничної та множинної імпутації. Для одиничної імпутації, в свою чергу, існують явний та неявний підходи моделювання (див. рис. 2). Явне моделювання, у випадку, коли прогнозований розподіл базується на формальній статистичній моделі. Неявне моделювання – коли фокус зосереджений на алгоритмі, який передбачає базову модель. Одинична імпутація передбачає заміну пропущеного значення одним новим.

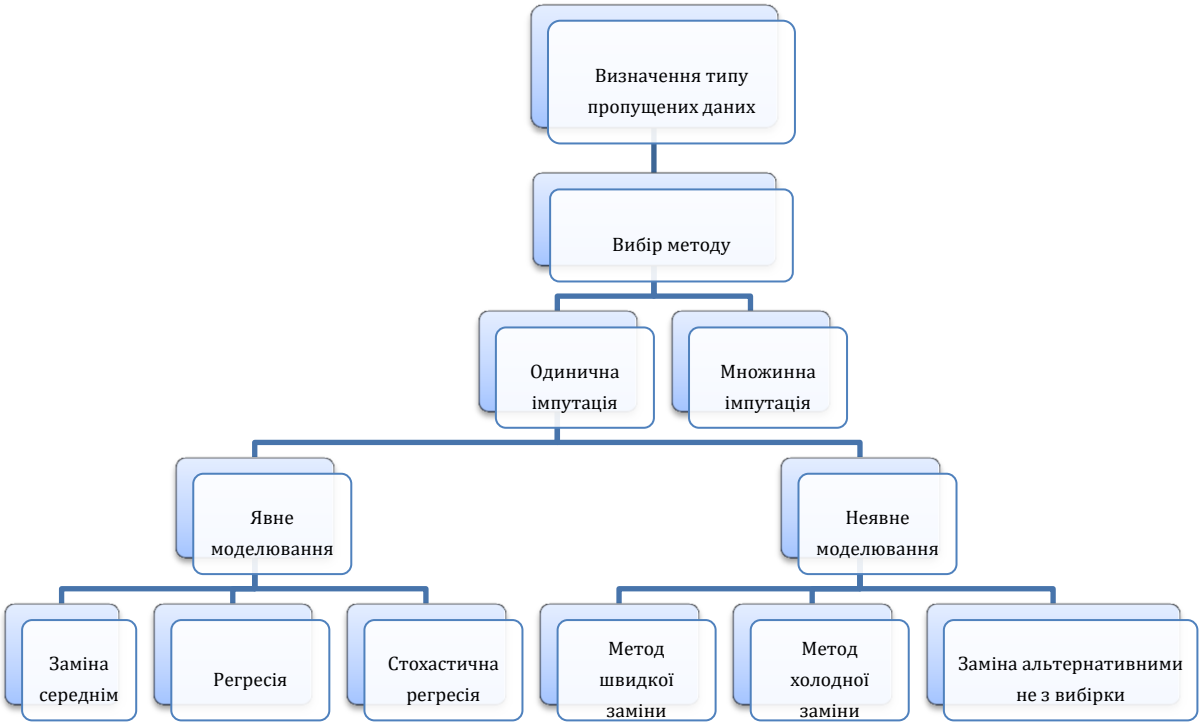

Рис. 2: Найбільш поширені методи

До найбільш поширених підходів явного моделювання належать: метод заміни середнім/модою, імпутація на основі регресивної моделі (відсутні значення замінюються новими прогнозованими з регресивної моделі) та імпутація на основі стохастичної регресивної моделі. До

основних підходів неявного моделювання пропущених значень належать метод швидкої заміни (англ. *not deck imputation*), заміна альтернативними значенням, взятими зі стороннього джерела зі схожою структурою (англ. *substitution*), метод холодної заміни (англ. *cold-deck imputation*) – полягає в заповненні пропусків константами зі стороннього ресурсу.

Відзначимо, що існують також комбіновані методи. Наприклад, метод швидкої заміни та регресії можна поєднати шляхом обчислення прогнозованих середніх на основі регресії.

На даний момент найкращою, і в той же час найбільш затратною в плані необхідних для обчислення ресурсів, є множинна імпутація (англ. *multiple imputation*). Даний метод передбачає заповнення пропусків масиву шляхом генерування декількох значень, одержаних в наслідок аналізу структури та з'язків між даними. Генерація кількох потенційних значень для кожного з пропусків покращує точність заміни і дозволяє певним чином оцінити якість імпутації (чим більша кількість значень, тим вища точність).

Найпоширенішим способом вирішення проблеми є усунення неповних спостережень лише у випадку коли для дослідження потрібна відсутнє в комірці значення. Якщо десь у наборі даних відсутні дані, у статистичному тестуванні використовуються існуючі значення. Оскільки попарне видалення використовує всю спостережувану варто зазначити низку недоліків:

1. Параметри моделі будуть залежати від різних наборів даних з різною статистикою (різний розмір вибірки, різні середні, різні дисперсії)
2. Такий спосіб зовсім не підходить у випадку не випадково відсутніх даних.

В нашій роботі використано п'ять різних методів, розглянемо більш детально їхню теоретичну основу.

2.1 Метод заміни середнім

Як зрозуміло з самої назви, даний метод полягає в заповненні пропусків масиву середнім по відповідній змінній. Інколи дещо кращий результат можна отримати при підстановці медіани або середнього зваженого. Найпростіший підхід не вимагає великих обчислювальних ресурсів, однак має значні недоліки, оскільки така підстановка призводить до зменшення дисперсії вибірки.

Рис. 3: Імпутація середнім

На прикладі значень CAC – головного показника французької фондової біржі за 1992-1998 роки – продемонстровано застосування методу до бази, з якої повністю випадковим чином було вилучено 5% даних (див. рис. 3). Дисперсія початкової вибірки рівна 336 764.6, дисперсія отриманої після підстановки – 312 709.5.

2.2 Метод швидкої заміни

Метод швидкої заміни полягає в заміщенні відсутніх даних відомими значення, донорами, взятими за схожого набору даних. В ідеалі кожен для кожного пропуску можна підібрати декілька донорів. Дослівно термін «not deck» перекладається як «гаряча колода», назва методу сягає середини XX століття, часу використання комп'ютерних перфокарт. Колоду (стопку) називали гарячою, оскільки, для заповнення пропусків використовувались перфокарти, що оброблялись на даний момент (на противагу методу холодного набору, коли в якості донорів брали перфокарти з інших колод).

Метод був розроблений Бюро перепису населення США. Щомісячно бюро проводить поточне дослідження населення (англ. The Current Population Survey, CPS) для збору інформації, яка стосується зайнятості населення та структури ринку праці, щомісячно проводиться анкетування щодо надбавки заробітку. Згідно статистики під час кожного описування близько 11-12% респондентів не даються відповіді на дане питання, і тоді для заповнення пропусків застосовують not deck. Суть полягає в заміщенні пропущеного значення показником запозиченим від іншого, якомога більш схожого за іншими параметрами, опитуваного. Замість того щоб замінювати відсутнє значення просто середнім чи випадковим значенням з масиву проводиться сортування даних на основі віку, статі, раси, роду зайнятості, посади, сімейного стану або інших показників. Записи сортуються на за місцем проживання опитуваних, а потім значення від респондентів використовуються для послідовного

заповнення відсутніх даних. Якщо для респондента не вдається знайти повну відповідність, not deck шукає спів падання з меншою кількість деталізації, прибираючи частину категорій.

Метод також активно застосовується в епідеміологічних і медичних дослідженнях. Hot deck забезпечує гнучкий підхід до роботи з відсутніми даними, що дає змогу зберігати внутрішню структуру і зв'язки без явних параметричних припущень моделі.

Розглянемо детальніше механізм роботи методу швидкої заміни.

Загальна ідея полягає у виборі одиниць донорів, які близькі до респондента в розумінні певної відстані. Нехай $x_i = (x_{i1}, \dots, x_{ik})$ є вектором k значень деякого спостереження, які використовуються для виявлення подібних класів, а $C(x_i)$ позначає клас в перехресній класифікації, до якої потрапляє i-те спостереження. Тоді пошук реципієнтів i та донорів j зводиться до пошуку мінімальної відстані, де метрику можна визначити як

$$d(i,j) = \begin{cases} 0, & j \in C(x_i) \\ 1, & j \notin C(x_i) \end{cases}$$

Групування даних на основі якісних характеристик – це не єдиний спосіб визначення подібних класів, можна визначити інші способи вимірювання відстані між потенційними донорами та реципієнтами. Маючи справу з неперервними величинами відстань можна визначити в розумінні відстані Махаланобіса

$$d(i,j) = \sqrt{(x_{i1} - x_{ik})^T S^{-1} (x_{i1} - x_{ik})}$$

Де S^{-1} – оцінка коваріаційної матриці x_i

Ще одним з варіантом є прогнозоване середнє зіставлення (англ. *predictive mean matching*), що обраховується на базі метрики

$$d(i,j) = \left(\hat{y}(x_i) - \hat{y}(x_j)\right)^2$$

де $\hat{y}(x_i)$ – прогнозоване значення Y на основі регресії, обчислене з використанням повних класів даних.

Існує два поширені способи застосування ход-дек імпутації [7]:

1. Метод випадкової швидкої заміни
2. Метод послідовної швидкої заміни

Другий підхід відрізняється від першого тим, що при визначенні донора ми відсортовуємо дані за однією зі змінних, а не обираємо будь-яке значення з групи випадковим чином. Відсутнє значення замінює донор, взятий з даних попереднього або наступного періодів. Використовувана для прогнозу величина – це змінна, яка, пов'язана зі

змінною, яка має бути приписана. Параметр для сортування обирається на основі наявності чи відсутності зв'язку між змінними.

Імпутація реалізується в два етапи. Перший етап полягає в кластеризації даних, а другий - в передачі значень, які вважаються такими, що мають подібні елементи. Формування кластерів здійснюється шляхом сортування даних за допомогою змінних предиктора. Потім, після сортування даних, відсутнє значення буде враховано за допомогою значення донора. Значення донорів отримують на основі змінних, які мають подібності в кластері, або з використанням даних попереднього чи наступного періодів. Недоліком методу є той факт, коли кількість відсутніх значень є достатньо великою, значення буде заповнюватися неодноразово, що призведе до зміщення оцінки.

Метод може бути заснований для даних, в яких пропущені значення розподілені рівномірно. Важливо врахувати залежність від спостережуваних змінних та зберегти внутрішню структуру між відсутніми і наявними значеннями. Метод краще підходить для застосування на великих масивах з великою кількістю показників, на основі яких можна проводити класифікацію.

2.3 Метод *k* найближчих сусідів

Метод *k* найближчих сусідів (англ. *k nearest neighbours, kNN*) може бути використаний для класифікації даних або відновлення пропущених значень. Застосовуючи kNN алгоритм для прогнозування відсутніх значень, ми беремо середнє або медіану, але не від всього масиву, а лише *k* «найближчих» значень до кожної пропущеної точки. Цей метод схожий на вищерозглянутий not deck, оскільки теж опирається на донорське запозичення значень. Перш за все алгоритм рахує відстані між всіма елементами масиву, на основі чого обираються k «найближчих» сусідів.

Для вимірювання відстані між сусідами можна використовувати декілька метрик, наприклад евклідову відстань чи відстань Махаланобіса (про які ми вже згадували вище). Якщо пропущені значення мають неперервний розподіл, то зазвичай застосовується евклідова відстань. Відстань Геммінга можна застосувати для прогнозування категоріальних даних. Відстань Геммінга — це кількість бітових позицій, у яких два біти відрізняються при порівнянні двох векторів однакової довжини, саме ця відстань є метрикою для порівняння порівняння векторів.

Найчастіше для вимірювання відстані між двома спостереженнями використовують відстань Гувера

$$d_{i,j} = \frac{\sum_{k=1}^{p} w_k \, \delta_{i,j,k}}{\sum_{k=1}^{p} w_k}$$

де w_k – зважені середні, а $\delta_{i,j,k}$ – внесок k-ї змінної.

Таким чином можна визначити відстані між неперервними, категоріальними та двійковими змінними, спосіб розрахунку $\delta_{i,j,k}$ залежить від виду даних. Для неперервних

$$\delta_{i,j,k} = \frac{|x_{i,k} - x_{j,k}|}{r_k}$$

Для категоріальних і двійкових

$$\delta_{i,j,k} = \begin{cases} 0, & x_{i,k} = x_{j,k} \\ 1, & x_{i,k} \neq x_{j,k} \end{cases}$$

Наступним кроком після знаходження k «сусідів» є імпутація пропущеного значення. Зазвичай беруть медіану цих значень, рідше середнє або арифметичне середнє. Алгоритм даного методу зображено на рис. 4.

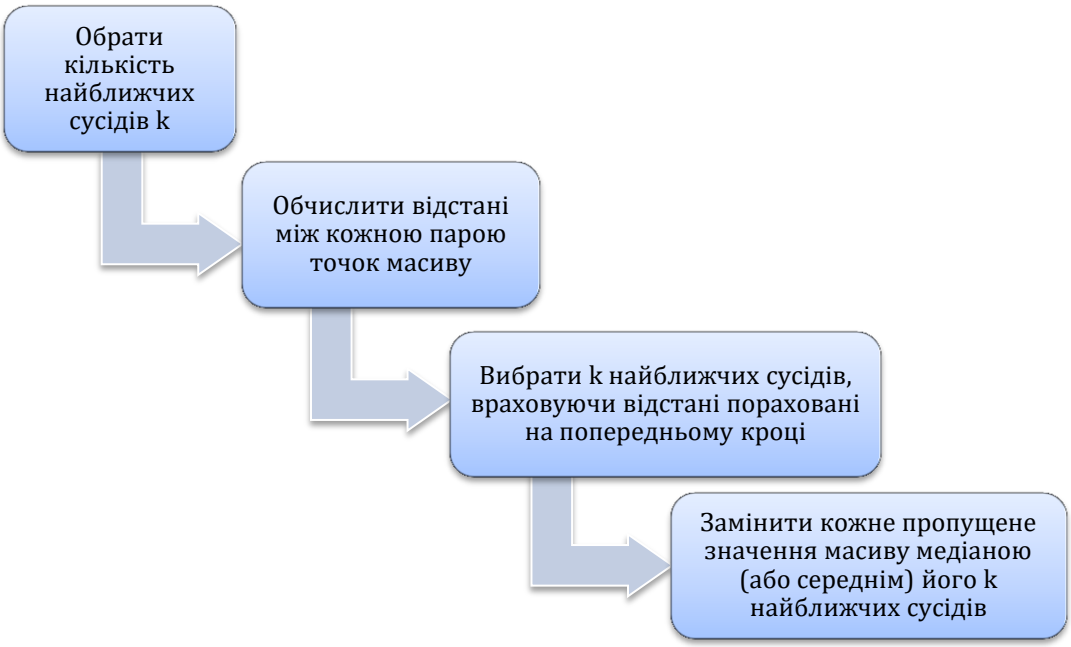

Рис. 4: Алгоритм kNN

Важливим питанням, від якого залежить якість імпутації, є вибір параметра k. Збільшення k може покращити заміну даних, просте надто велике значення параметру порушить ідею kNN методу обирати донора з невеликої групи сусідів, а не всього масиву (в крайньому випадку, якщо k рівне кількості спостережень, підхід зведеться до заміни всіх пропусків на середнє/медіану).

Очевидно, що підібрати k потрібно так аби імпутовані змінні були якомога ближчими до пропущених, проте не маючи повного набору встановити напевно значення k неможливо. Одна з ідей підбору полягає в застосуванні перехресної перевірки. Необхідно знайти середнє (або медіану) по кожній змінній, що містить пропущені значення та визначити приблизні межі параметру k, після чого замінити всі пропущені значення на середнє (або медіану) і вибрати таке k, при якому абсолютна середня різниця між значеннями імпутованими за допомогою методу заміни середнім/медіаною та методом k найближчих сусідів буде мінімальною. Аналітично визначити найкраще k дуже складно, проте нижче наведено схему підбору параметру, що ґрунтується на принципі машинного навчання перехресний підбір (англ. *cross-validation*) [11].

Принцип підбору параметру k

1. **Вхідні дані:** кількість пропусків n, межі параметру $[k_1;k_2]$

2. Заповнення всіх пропусків середнім/медіаною.

3. Створення порожнього списку ERROR довжини k_1-k_2

4. **Для всіх i в межах $[k_1; k_2]$** застосувати метод kNN(), порахувати середню різницю ERROR між значеннями заповненими середніми та за допомогою kNN

$$ERROR_i = \frac{\sum_{j=1}^{n} ERROR_{i,j}}{n}$$

де $i = k_1, k_1 + 1, \dots, k_2 - 1, k_2$

5. Обрати таке значення параметру, при якому середня різниця буде мінімальною

$$k_{\text{найкр}} = \underset{i}{\operatorname{argmin}}\, ERROR_i$$

6. Вивести масив заповнений при $k = k_{\text{найкр}}$

Зауважимо, що наведений алгоритм, як і сам метод k найближчих сусідів, на відміну від двох розглянутих раніше способів, досить ресурсозатратний і підбір найкращого параметру таким чином часто буде недоцільним чи навіть неможливим.

Крім того, наведений алгоритм підбору параметру не обов'язково приведе нас до найбільш підходящого значення як ми пізніше переконаємось на прикладі.

2.4 Метод випадкового лісу

В попередніх розділах ми розглянули деякі прості методи як заміна середнім/медіаною чи метод швидкої заміни. Хоча вони легкі для розуміння і їх просто програмувати, вони часто не працюють належним чином, адже використовуючи їх ми недооцінюємо дисперсію або робимо велику кількість значень одноковими, що часто зовсім не відповідає дійсності. В цьому розділі ми звернемо увагу на більш складний і ефективний метод.

2.4.1 Історія виникнення

Алгоритм випадкового лісу (англ. *random forest, RF*) був останньою працею американського статистика Лео Бреймана (рис. 5), Брейман, математик Каліфорнійського університету (широко відомого як Берклі), в 2001 році розробив його для покращення класифікації даних. Його робота, в свою чергу, була натхненна ідеєю опублікованою в 1997 році в Ялі Амітом та Дональдом Германом в публікації під назвою «Квантування та розпізнавання за допомогою випадкових дерев».

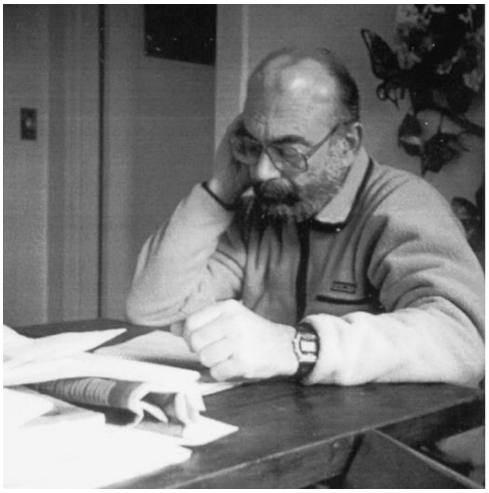

Рис. 5: Лео Брейман (1928-2005)

Метод був розроблений для аналізу даних у Weka (скорочено від *Waikato Environment for Knowledge Analysis*) – програмному забезпеченні для аналізу даних, яке розробляє Університет Вайкато (англ. *University of Waikato*). Випадковий ліс може бути застосований для кластеризації даних і заповнення пропущених значень в базі даних. Однією з переваг методу є те, що предиктори можуть бути як категоріальними, так і неперервними.

Рис. 6: Деніел Стекховен

В січні 2012 року в журналі «*Genomics, Proteomics & Bioinformatics*» швейцарськими математиками Деніелом Стекховеном (рис. 6) і Пітером Бюльманом (рис. 7) була опублікована стаття під назвою «MissForest — непараметрична імпутація пропущених значень для даних змішаного типу» в якій автори запропонували спосіб використання алгоритму випадкового лісу для вирішення задачі пропущених даних, використовуючи ітераційну схему імпутації.

Рис. 7: Пітер Бюльман (1965)

Описаний в науковій праці метод підходить для імпутації даних для змінних мішаного типу та дуже добре працює в «поганих» умовах, як, наприклад, великі розміри масиву чи складні взаємозв'язки. Завдяки своїй точності та надійності підхід, в основі якого лежить алгоритм

випадкового лісу зручний для використання в прикладних дослідженнях, особливо сфері медичної статистики.

2.4.2 Random forest

Випадковим лісом називають ансамбль дерев рішень, де кожне дерево залежить від деякого набору випадкових змінних.

Нехай вектор $X = (X_1, .., X_p)^T$ являє собою набір предикторів, а Y – залежна від X змінна, сумісний розподіл $P_{XY}(X, Y)$ нам невідомий. Задача полягає у встановленні залежності $f(X) = Y$, а задача визначення функції f зводиться до задачі знаходження функції втрат $L(Y, f(X))$ і мінімізації її значення

$$E_{XY}(L(Y, f(X)))$$

Природнім вибором в якості функції L було би взяти середньоквадратичну похибку для неперервних даних

$$L(Y, f(X)) = (Y - f(X))^2$$

або

$$L(Y, f(X)) = \begin{cases} 0, & Y = f(X) \\ 1, & Y \neq f(X) \end{cases}$$

для категоріальних даних.

Потреба мінімізації $L(Y, f(X))$ приводить до регресивної функції

$$f(x) = E(Y|\mathrm{X} = \mathrm{x})$$

У випадку дискретних змінних маємо правило Байєса

$$f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} P(Y = y|X = x)$$

де $\mathcal{Y}$ – множини всіх можливих значень функції.

Таким чином, сконструювати функцію предиктор f можна на основі так званих базових учнів (англ. *base learners*) $h_1(x), \dots, h_J(x)$. Для випадку регресії маємо

$$f(x) = \frac{\sum_{j=1}^{J} h_j(x)}{J}$$

для класифікації

$$f(x) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{J} I(y = h_j(x))$$

де

$$I\big(Y \neq f(X)\big) = \begin{cases} 0, & Y = f(X) \\ 1, & Y \neq f(X) \end{cases}$$

Для розуміння принципу роботи алгоритму випадкового лісу, потрібно мати деякі базові знання про тип дерев, які використовуються як базові учні. Дерева, що застосовуються в даному алгоритмі утворюються на основі двійкового рекурсивного принципу, вони розділяють предиктор використовуючи послідовності окремих змінних.

Кореневий вузол дерева включає в себе весь простір предиктора. Нерозділені вузли називають кінцевими вузлами, вони формують остаточний простір предиктора. Кожен не кінцевий вузол розбивається на два дочірні вузли, лівий і правий, відповідно до значення однієї зі змінних предиктора.

Для неперервних змінних розбиття відбувається за принципом «менше-більше». Менше стає лівим нащадком, більше – правим. Для категоріальних змінних лівим нащадком стає змінна, яка не міститься в множині значень, правим – та, яка належить множині (рис. 8).

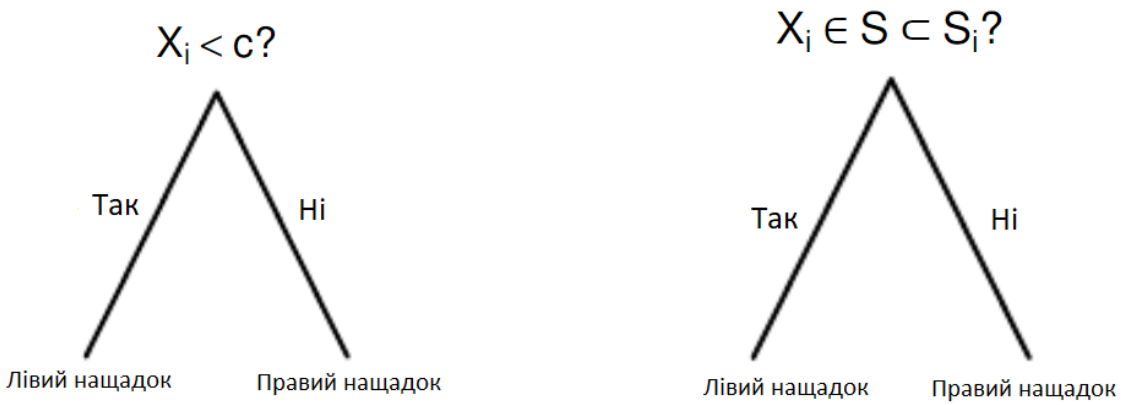

Рис. 8: Формування нащадків для випадку неперервних змінних (зліва) та категоріальних (справа)

Конкретне розбиття, яке дерево використовує для поділу, обирається шляхом розгляду кожного можливого розбиття на кожній змінній предиктора та вибору найбільш відповідного згідно певного критерію. У випадку регресії, для значень вузла $y_1, \ldots, y_n$, природнім буде розщеплення є середньоквадратичне залишкове відхилення

$$Q = \frac{\sum_{i=1}^{n}(y_i - \hat{y_i})^2}{n}$$

де

$$\hat{y_i} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

У випадку класифікації, де K класів позначимо як $1, \dots, K$ найбільш вживаним критерієм буде індекс Джинні

$$Q = \sum_{k \neq k'}^{K} \hat{\rho}_k \, \hat{\rho}_{k'}$$

де $\hat{\rho}_k$ – це пропорція спостережень класу у вузлі

$$\hat{\rho}_k = \frac{\sum_{i=1}^{n} I(y_i = k)}{n}$$

Процедура продовжується рекурсивно, до тих пір поки не буде виконано критерій зупинки, часто в програмуванні наперед встановлюється певне обмеження кількості нащадків.

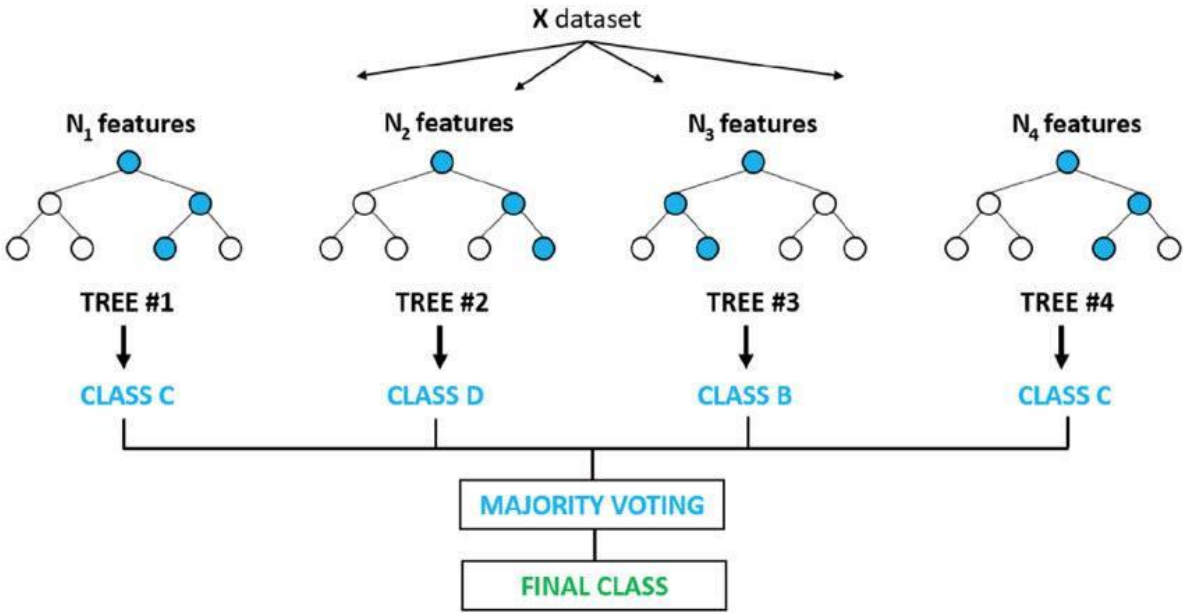

Рис. 9: Схема роботи алгоритму випадкового лісу

Прогнозоване значення обчислюється шляхом знаходження середнього в задачі регресії або виявленню найпоширенішого класу в задачі класифікації. Щоб спрогнозувати точку, набір значень її предиктора використовується для передачі точки вниз по дереву, доки вона не потрапить у кінцевий вузол, а передбачення для кінцевого вузла використовується як передбачення для нової точки.

Схема роботи алгоритму зображена на рис. 9.

2.4.3 MissForest

Перша за все, звернемо увагу на те, що під терміном Random Forest можуть розуміти як алгоритм класифікації даних так і метод імпутації, що базується на даному алгоритмі. Термін MissForest використовують лише для опису методу імпутації вперше описаного Стекховеном і Бюльманом в 2012 році. Саме MissForest алгоритм ми застосовуємо в даній дисертації.

MissForest — це алгоритм машинного навчання, який дозволяє обробляти категоріальні та кількісні змінні, не роблячи конкретних припущень щодо структури та розподілу даних.

На першому кроці алгоритм замінює всі відсутні значення середнім/медіаною для неперервних змінних та значенням, що найчастіше зустрічається, для категоріальних. Далі послідовно, починаючи зі змінної, яка містить мінімальну кількість пропущених значень, виконується процедура імпутації. Алгоритм випадкового лісу навчається на все більш якісних даних на кожному кроці ітерації. Тренування відбувається на спостереженнях, що містять повну інформацію і використовується для прогнозування відсутніх даних.

Ітераційний процес триває поки не буде виконано критерій зупинки або не буде досягнуто деякої заданої наперед кількості ітерацій.

Опишемо роботу алгоритму більш формально.

Нехай X – матриця розмірності $n \times p$, де n – кількість спостережень, а p – кількість змінних. Не обов'язково брати до уваги всі змінні масиву, потрібні лише ті, які можуть бути використані для імпутації пропущених даних. Позначимо всі записи, що містять відсутні дані як $Y_k,\ k = 1, \dots, p$.

Тоді можемо умовно розділити масив на частини:

- y_k^{obs} – вектор наявних значень Y_k
- y_k^{mis} – вектор пропущених значень Y_k
- X_k^{obs} – матриця змінних $X_i,\ i = 1, \dots, p$ без пропущених даних в значеннях Y_k

- x_k^{mis} – матриця змінних X_i, $i = 1, \dots, p$ з пропущеними даними в значеннях Y_k

X_1	X_2	…	X_p	X_k
x_{11}^{obs}	x_{12}^{obs}	…	x_{1p}^{obs}	y_{1k}^{obs}
x_{21}^{obs}	x_{22}^{obs}	…	x_{2p}^{obs}	…
…	…	…	…	…
$x_{n_1 1}^{obs}$	$x_{n_1 2}^{obs}$	…	$x_{n_1 p}^{obs}$	$y_{1n_1}^{obs}$
x_{11}^{mis}	x_{12}^{mis}	…	x_{1p}^{mis}	y_{1k}^{mis}
…	…	…	…	…
$x_{n_2 1}^{mis}$	$x_{n_2 2}^{mis}$	…	$x_{n_2 p}^{mis}$	$y_{1n_2}^{mis}$

Рис. 10: Матриця розміру $\boldsymbol{n \times p}$ з пропущеними значеннями, $\boldsymbol{n_1 + n_2 = n}$

Суть методу полягає в тому, щоб знайти таку функцію f, при якій виконуватиметься

$$\left(y_k^{obs} - f(x_k^{obs})\right)^2 = 0$$

і таким чином мати змогу y_k^{mis} за допомогою $f(x_k^{mis})$.

MissForest

1. **Вхідні дані:** матриця X розмірності $n \times p$ та критерій зупинки γ.

2. Заповнюємо всі пропуски середнім/медіаною по відповідній змінній.

3. $k \longleftarrow$ вектор відсортованих згідно збільшення кількості пропущених значень колонок масиву X

4. Поки не виконана умова $\boldsymbol{\gamma}$:

$\mathrm{X}_{\mathrm{old}}^{\mathrm{imp}} \longleftarrow$ зберігаємо отриманий на попередьому кроці масив

for s in k:

застосувати метод випадкового лісу $y_{obs}^s \sim x_{obs}^s$

провести імпутацію y_{mis}^s на основі x_{mis}^s

$\mathrm{X}_{\mathrm{new}}^{\mathrm{imp}}$ – вставити в масив y_{mis}^s знайдені на s-му кроці ітерації

5. Вивести масив $\mathrm{X}^{\mathrm{imp}}$

Умова зупинки циклу γ залежить від виду змінної з пропусками. Ітерації припиняються як тільки вперше зростає різниця.

Для неперервних даних

$$\Delta N = \frac{\sum_{j \in N} (\mathrm{X}_{\mathrm{old}}^{\mathrm{imp}} - \mathrm{X}_{\mathrm{new}}^{\mathrm{imp}})^2}{\sum_{j \in N} (\mathrm{X}_{\mathrm{new}}^{\mathrm{imp}})^2}$$

Для категоріальних даних

$$\Delta F = \frac{\sum_{j \in F} \sum_{i=1}^{n} 1_{\mathrm{X}_{\mathrm{old}}^{\mathrm{imp}} \neq \mathrm{X}_{\mathrm{new}}^{\mathrm{imp}}}}{n_{NA}}$$

де n_{NA} – загальна кількість пропусків серед категоріальних змінних.

Якщо в масиві були пропуски серед змінних обох видів, то алгоритм зупиняється коли ці дві умови виконуються одночасно.

Розглянутий нами метод має багато переваг. Алгоритм missForest дає змогу проводити імпутацію майже будь-яких даних, в тому числі змішаних коли присутні і неперервні і категоріальні змінні. Нам не потрібно робити припущення про розподіл даних або наперед підбирати параметри. Все це робить метод випадкового лісу дуже популярним, він широко використовується в багатьох сферах, зокрема медицині, банківському секторі та маркетингу.

2.5 Метод максимального градієнтного підсилення

Градієнтне підсилення — це метод машинного навчання, який послідовно поєднує і аналізує дерева рішень. Він спрямований на покращення загальної ефективності прогнозування шляхом оптимізації вагових коефіцієнтів моделі на основі помилок в попередніх ітераціях та підвищення точності моделі на кожному кроці.

Метод максимального градієнтного підсилення для імпутації пропущених значень (англ. *eXtreme Gradient Boosting, XGBoost*) вперше був опублікований математиками Йонші Денгом та Томасом Ламлі в Журналі обчислювальної та графічної статистики 19 жовтня 2023 року в статті під назвою «Множинна імпутація за допомогою XGBoost» [14].

XGBoost як і алгоритм випадкового лісу — це метод ансамблювання, який може класифікувати дані шляхом комбінування результатів окремих дерев, просте ці алгоритми відрізняються одне від одного способом побудови дерев і способом об’єднання результатів.

Як і в попередньому розділі, ми розглянемо загальну модель роботи алгоритму і його застосування для вирішення проблеми пропущених даних.

2.5.1 XGBoost model

Дерево рішень для XGBoost алгоритму – це модель, де кожен вузол, являє собою розбиття на конкретну функцію, кожна гілка представляє потік даних на основі розбиття, а кожен лист представляє класифікацію. Дерево рішень формується завдяки упорядкуванню функцій і перевірки кожного можливого розподілу для кожної функції. Для кожного вузла поділ на нащадків обирається такий, який дає найбільш підходящий результат для деякої функції.

Процес триває доти, доки не буде виконана умова зупинки, зазвичай це відбувається тоді коли алгоритм досягає максимальної кількості ітерацій або спрацьовує правильно дострокового припинення алгоритму.

Нарешті, щоб зробити прогноз, кожен лист повинен мати асоційоване з ним значення. Відповідь (англ. *response*) листа зазвичай буде відповіддю більшості його навчальних прикладів для проблем класифікації чи середньої відповіді навчальних прикладів для проблем регресії.

В цілому опис роботи алгоритму нагадує розглянутий в попередньому розділі спосіб тренування і вибору прогнозу, але на відміну від методу випадкового лісу дана модель будується на основі процесу бустингу (англ. *boosting*), не беггінгу (англ. *begging*).

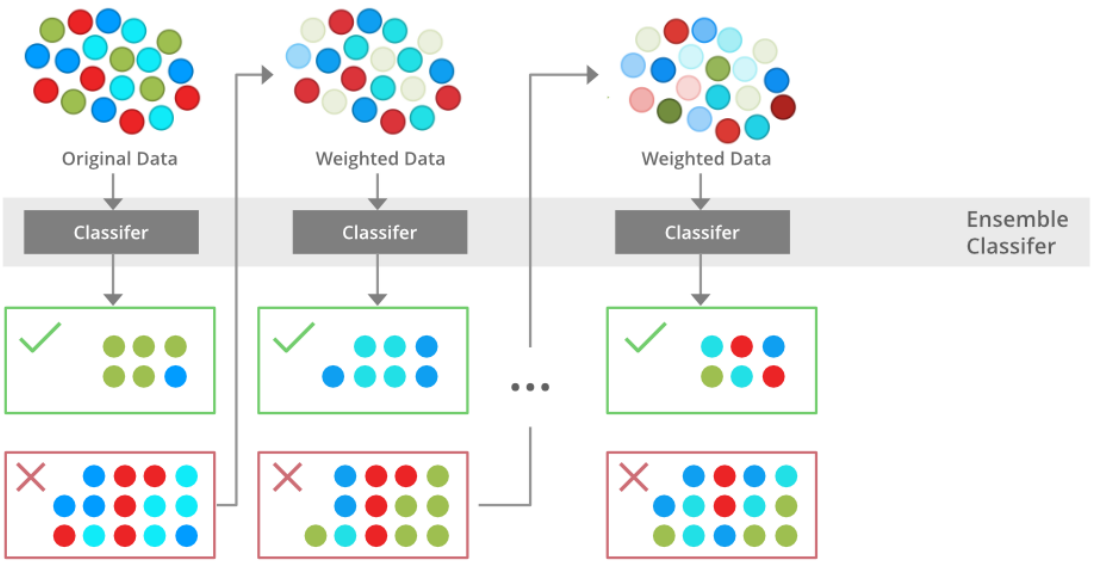

Рис. 11: Схема роботи алгоритму максимального градієнтного підсилення

Різниця полягає у способі навчання слабких учнів (англ. *weak learners*), бустингу передбачає вибір деякого набору даних і присвоєння однакової ваги для кожної точки набору – це вхідні дані моделі за допомогою, який визначаються неправильно класифіковані точки даних. Потім ми збільшуємо вагу неправильно класифікованих точок даних і зменшуємо вагу класифікованих правильно, а потім нормалізуємо вагу всіх точок, процес триває поки не буде досягнуто потрібного результату. Схема роботи алгоритму зображена на рис. 11, її можна візуально порівняти зі алгоритмом беггінгу схематично зображеному на рис. 9.

Нехай D – масив даних розміром $n \times m$, де n – кількість спостережень, а m – кількість змінних, що буде використовуватись для прогнозування. Модель складається з K адитивних функції [15]

$$\hat{y}_i = \varphi(x_i) = \sum_{k=1}^{K} f_k(x_i)$$

де функції f_k являють собою простій всіх регресивних дерев ухвалення рішень.

$$\mathcal{F} = \{f(x) = w_{q(x)}\}(q: \mathbb{R}^m \to T, w \in \mathbb{R}^T)$$

де T – кількість листів дерева

Щоб визначити ці функції ми повинні мінімізувати наступну функцію

$$\mathcal{L}(\varphi) = \sum_i l(\hat{y}_i; y_i) + \sum_k \Omega(f_k)$$

де

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$$

де l – диференційована опукла функція втрат, а Ω відображає складність моделі, тобто функції дерева регресії. Додатковий член регулярізації допомагає згладити остаточні вивчені ваги, щоб уникнути надмірної підгонки.

Нехай $\hat{y}_i^{(t)}$ – прогноз на t-й ітерації. Потрібно підібрати f_t так щоб мінімізувати наступний вираз

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i{}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

Наближення другого порядку використовується для швидкої оптимізації в загальному випадку (для спрощення моделі приберемо константи)

$$\hat{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \Omega(f_t)$$

де

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i; \hat{y}^{(t-1)})$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i; \hat{y}^{(t-1)})$$

Поклавши $I_j = \{i, q(x_i) = j\}$ можна переписати $\hat{\mathcal{L}}^{(t)}$ у вигляді

$$\hat{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} g_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 =$$

$$= \sum_{i=1}^{n} \left[w_j \sum_{i \in I_j} g_i + \frac{1}{2} w_j^2 \sum_{i \in I_j} (h_i + \lambda) \right] + \gamma T$$

Для деякої фіксованої структури дерева $q(x)$ можна виразити оптимальну вагу як

$$w_j^2 = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} (h_i + \lambda)}$$

тоді відповідне оптимальне значення функції втрат матиме вигляд

$$\hat{\mathcal{L}}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} (h_i + \lambda)} + \gamma T$$

Остання формула може бути використана для якісної оцінки структури дерева прийняття рішень $q(x)$.

Нехай I_L та I_R – множини, що складаються з лівих і правих нащадків які утворюються після ітерації.

$$I = I_L \cup I_R$$

Тоді $\mathcal{L}$ можна виразити наступним чином

$$\mathcal{L}_{split} = \left[\frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in L}(h_i + \lambda)} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R}(h_i + \lambda)} - \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j}(h_i + \lambda)} \right] - \gamma$$

На практиці, як правило, застосовують саме таку формулу.

2.5.2 XGBoost imputation

Метод множинної імпутації був вперше запропонований Рубіном в його роботі «Множинна імпутація пропущених даних при опитуваннях», опублікованій в 1978 році. З часом підхід ставав все більш і більш популярним методом імпутації, оскільки давав змогу зменшити зміщення розподілу та відобразити невизначеність відсутніх значень.

Суть полягає в тому щоб для кожного з пропущених значень підібрати набір подібних донорі, проаналізувати отримані множини і обрати найбільш підходящу заміну для імпутації. Рубін стверджував, що за належної багаторазової імпутації об'єднані оцінки та дисперсія будуть статистично достовірними.

В дисертації використовується метод максимального градієнтного підсилення для імпутації пропущених значень, що був опублікований Денгом та Ламлі в 2023 році, його повну назву можна перекласти як «множинна імпутація за допомогою методу максимального градієнтного підсилення з використанням зіставлення прогнозованого середнього та підвибірок» [14].

В машинному навчанні підвибіркою (англ. *subsampling*) називають практику випадкового вибору підмножини точок із більшого набору даних для тренувальних моделей. Підмножина вказується шляхом вибору параметра n, вказуючи, що кожна *n*-та точка вилучена. Це схоже на вибір невеликої «розумної» групи даних замість використання всього набору, може бути корисним для заощадження оперативної пам'яті.

Нехай матриця Y_{raw} містить p змінних з пропущеними значеннями, першим кроком виконання алгоритму буде сортування змінних в порядку зростання пропущених значень в кожній з них. Після цього замінюємо всі

пропуски в масиві середніми по відповідних змінних (або робимо перше припущення користуючись певними міркуваннями) – отримуємо повністю заповнену матрицю Y_{init}.

Позначимо через Y_i^{obs} та Y_i^{mis} відповідно наявні та пропущені значення змінної Y_i, також позначимо через Y_{-i}^{obs} та Y_{-i}^{mis} всі дані у всіх змінних, де зустрічаються значення Y_i, крім Y_i.

Для кожної змінної з пропущеними значеннями, ми застосовуємо модель XGBoost з підвибірками для отримання прогнозів щодо Y_i^{mis} на основі Y_i^{obs}.

Ще Рубін в 1986 році зауважив, що для неперервних змінних імпутація пропущених значень за допомогою точкових предикторів $\hat{Y}_i^{*mis}$ може призвести до недооцінки варіативності імпутації, саме для вирішення цієї проблеми може бути корисним PMM метод.

Зіставлення прогнозованого середнього (англ. *predictive mean matching, PMM*) обчислює прогнозоване значення змінної згідно деякої схеми імпутації, як правило шляхом зіставлення прогнозованого значення кожного відсутнього запису з набором донорів, що мають найближчі прогнозовані значення серед наявних повних записів. Для кожного відсутнього значення створюється множина з потенційних донорів. Випадковим чином обирається елемент з множини донорів та ампутується замість пропущеного значення цільової змінної.

Нехай $\hat{\beta}_i$ оцінка моделі $f: Y_i^{obs} \longrightarrow Y_{-i}^{obs}$ з використанням всього набору даних і $\hat{\beta}_i^*$ оцінка моделі $f^*: Y_i^{*obs} \longrightarrow Y_{-i}^{*obs}$ з використанням PMM при кожній імпутації.

Донори	Реципієнти	Відмінності
$\hat{Y}_i^{obs} = Y_{-i}^{obs}\hat{\beta}_i$	$\hat{Y}_i^{mis} = Y_{-i}^{mis}\hat{\beta}_i$	$\hat{\beta}_i$ однакові для кожної імпутації
$\hat{Y}_i^{obs} = Y_{-i}^{obs}\hat{\beta}_i$	$\check{Y}_i^{*mis} = Y_{-i}^{mis}\hat{\beta}_i^*$	Донори $\hat{\beta}_i$ використовують реципієнтів $\hat{\beta}_i^*$ для всіх імпутацій
$\check{Y}_i^{*obs} = Y_{-i}^{obs}\hat{\beta}_i^*$	$\check{Y}_i^{*mis} = Y_{-i}^{mis}\hat{\beta}_i^*$	Донори і реципієнти використовують $\hat{\beta}_i^*$ для кожної імпутації
$\check{Y}_i^{\prime obs} = Y_{-i}^{obs}\hat{\beta}_i^{\prime}$	$\check{Y}_i^{\prime\prime mis} = Y_{-i}^{mis}\hat{\beta}_i^{\prime\prime}$	$\hat{\beta}_i$ та $\hat{\beta}_i^*$ - різні оцінки для кожної імпутації

Рис. 12: способи застосування PMM методу

Чотири можливі способи застосування PMM наведено на рис. 12.

3 ОСНОВНА ЧАСТИНА

В попередніх розділах було розглянуто п’ять різних підходів до заміни пропущених даних:

1. Метод заміни середнім
2. Метод швидкої заміни
3. Метод k найближчих сусідів
4. Метод випадкового лісу
5. Метод максимального градієнтного підсилення

За допомогою цих методів ми проведемо імпутацію пропущених в масиві значень, в якості бази даних візьмемо дані з офіційного сайту Центру досліджень епідеміології катастроф (CRED), URL: https://public.emdat.be/data

DisNo.	Disaster _Grou	Disaster_Subgroup	Disaster_Type	Country	Subregion	Region	Date	Total_Affected	Total_Damage
1900-0003-USA	Natural	Meteorological	Storm	United States of	Northern Amer	Americas	08.09.1900	6 000,00	1 098 720,00
1902-0012-GTM	Natural	Geophysical	Earthquake	Guatemala	Latin America a	Americas	18.04.1902	2 000,00	880 384,00
1905-0003-IND	Natural	Geophysical	Earthquake	India	Southern Asia	Asia	04.04.1905	20 000,00	847 777,00
1905-0008-ALB	Natural	Geophysical	Earthquake	Albania	Southern Europ	Europe	01.06.1905	120,00	807 084,00
1906-0013-USA	Natural	Geophysical	Earthquake	United States of	Northern Amer	Americas	18.04.1906	700,00	17 769 416,00
1906-0014-CHL	Natural	Geophysical	Earthquake	Chile	Latin America a	Americas	16.08.1906	20 000,00	3 391 110,00
1906-0015-HKG	Natural	Meteorological	Storm	China, Hong Kon	Eastern Asia	Asia	08.09.1906	10 000,00	678 222,00
1907-0004-JAM	Natural	Geophysical	Earthquake	Jamaica	Latin America a	Americas	14.01.1907	1 200,00	981 000,00
1908-0007-ITA	Natural	Geophysical	Earthquake	Italy	Southern Europ	Europe	28.12.1908	75 000,00	3 933 688,00
1911-0006-TWN	Natural	Meteorological	Storm	Taiwan (Province	Eastern Asia	Asia	01.08.1911	1 000,00	654 000,00
1912-0008-JPN	Natural	Meteorological	Storm	Japan	Eastern Asia	Asia	01.09.1912	1 000,00	631 448,00
1912-0021-CAN	Natural	Meteorological	Storm	Canada	Northern Amer	Americas	30.06.1912	28,00	157 862,00
1913-0005-USA	Natural	Meteorological	Storm	United States of	Northern Amer	Americas	01.03.1913	732,00	6 156 540,00
1914-0002-JPN	Natural	Geophysical	Volcanic activity	Japan	Eastern Asia	Asia	03.01.1914	140,00	609 497,00
1915-0002-ITA	Natural	Geophysical	Earthquake	Italy	Southern Europ	Europe	13.01.1915	29 980,00	1 810 388,00
1915-0004-USA	Natural	Geophysical	Earthquake	United States of	Northern Amer	Americas	22.06.1915	6,00	30 173,00
1915-0005-USA	Natural	Meteorological	Storm	United States of	Northern Amer	Americas	01.08.1915	525,00	1 810 388,00
1916-0002-CAN	Technolc	Miscellaneous accid	Fire (Miscellane	Canada	Northern Amer	Americas	30.07.1916	228,00	69 896,00

Рис. 13: Частина бази з сайту CRED

В таблиці у стовпчику Total_Affected вказано загальну кількість постраждалих (загиблих), у стовпчику Total_Damage стоять суми завданих катастрофою збитків в тисячах американських доларів.

Для дослідження методів було обрано повні дані щодо природніх та техногенних катастроф останні за 124 роки. Ця база була обрана оскільки саме дані такого типу часто становлять велику проблему в статистичних дослідженнях. Щоб порахувати адекватну оцінку ризиків пов’язаних з природніми чи техногенними катастрофами потрібна якомога більш повна інформація про кількість постраждалих і суми завданих збитків, але дуже часто виникають проблеми через закритість інформації або неможливість виміряти завдану шкоду.

Кількість записів в таблиці становить 4399, ми випадковим чином видалимо 10% даних зі категорії Total_Damage та 5% з категорії Total_Affected. Такий характер пропусків буде природнім для такого виду даних, з іншими змінними, як правило, не виникає проблем (тобто час або локацію виникнення пожежі або землетрусу неважко встановити без застосування спеціальних методів, але суми завданої шкоди часто невідомі).

3.1 Застосування

В цьому розділі ми опишемо особливості застосування кожного з методів. Результат залежить від того які саме дані будуть пропущені в масиві, тому зафіксуємо set.seed(97) для можливості потім відтворити розподіл пропущених значень і результати застосування імпутації.

Метод заміни середнім надзвичайно простий в застосуванні. Знаходимо середні та медіани по змінним Total_Damage та Total_Affected і заміщуємо всі пропуски відповідними значеннями.

При застосуванні алгоритму hot deck обираємо донора для імпутації відсортувавши дані в кожній групі за змінною «Region».

Метод k найближчих сусідів дещо складніший в застосуванні ніж два попередні, адже вимагає наперед заданого параметра k. В розділі присвяченому теоретичному розгляду kNN алгоритму ми наводили принцип пошуку оптимального параметру, просте є суттєві перешкоди для втілення такої ідеї на практиці. Метод, що вимагає обчислення відстаней між всіма точками і так вимагає більше ресурсів ніж простіші підходи, додатковий пошук параметру робить програмне втілення принципу ще затратнім. Крім того, немає певності, що знаходження k, яке даватиме найменшу середню похибку по змінній з пропусками заповненими kNN алгоритмом і пропусками, заповненими за допомогою методу середнього буде давати найменшу похибку при порівнянні з істинними значеннями, а розв’язуючи більшість задач немає ніякої змоги отримати повний набір (власне імпутація і застосовується для заміни пропущеної інформацію на прогнозовані значення).

Зазвичай в якості k беруть невеликі значення близько десяти. Для нашого алгоритму k = 15. Також потрібно визначити як вибирається значення для імпутації після знаходження найближчих сусідів для кожного пропуску. Застосуємо метод використавши медіану.

Для методу випадкового лісу (а якщо точніше, то алгоритму під назвою MissForest) потрібно накласти обмеження на кількість ітерацій. Хоча для програми прописані умови зупинки циклу, варто розуміти, що час виконання інколи може бути неприйнятним, тому доцільно наперед

визначити максимальну кількість ітерацій, в нашому випадку maxiter = 10 (це найбільш поширене обмеження). Кількість дерев рівна 100. В якості початкового припущення взято середні значення по відповідних змінних.

Найскладнішим і найбільш ресурсозатратним алгоритмом з тих, що ми розглядали є метод максимального градієнтного підсилення. Це метод множинної імпутації, кількість ампутованих множин в нашому випадку рівна 4, кількість ітерації – 2 (цього цілком достатньо).

Особливістю даного варіанту алгоритму є використання методу зіставлення прогнозованого середнього для недопущення недооцінки варіативності імпутації для неперервних змінних. Змінні Total_Affected та Total_Damage є неперервними з точки зору заміщення пропущених в них значень, використовуємо вид CMM, що наведено нижче

Донори	Реципієнти	Відмінності
$\hat{Y}_i^{obs} = Y_{-i}^{obs}\hat{\beta}_i$	$\check{Y}_i^{*mis} = Y_{-i}^{mis}\hat{\beta}_i^*$	Донори $\hat{\beta}_i$ використовують реципієнтів $\hat{\beta}_i^*$ для всіх імпутацій

3.1 Якісна оцінка методів

Оскільки пропуски в масиві були створені штучно, є змога порівняти відхилення результатів одержаних за допомогою методів імпутації від істинних значень. В якості критеріїв для якісної оцінки оберемо нормалізовану середньоквадратичну похибку NRMSE (англ. *normalized root mean squared error*) та середню абсолютну похибку (англ. *mean absolute error*).

$$NRMSE = \frac{\sqrt{\sum_{i=1}^{n} \frac{(y_i' - y_i)^2}{n}}}{y_{max} - y_{min}}$$

$$MAE = \sum_{i=1}^{n} \frac{|y_i' - y_i|}{n}$$

де y – істині значення, y_i' - імпутовані значення, n – кількість записів в масиві.

Для оцінки роботи алгоритмів імпутації частіше використовується лише перша, проте враховуючи широкий діапазон значень по обох

змінних з пропущеними даними отримані результати можуть бути не дуже наочними, тому варто використати ще один спосіб порівняння.

Метод	NRMSE		MAE	
	Матеріальні збитки	Кількість постраждалих	Матеріальні збитки	Кількість постраждалих
Mean imputation	0.009335907	0.01507282	248363.3	902.8898
Median imputation	0.009535847	0.01507754	191715.7	854.1107
Hot deck	0.012849825	0.01511458	300080.9	925.6501
kNN	0.009171064	0.01507463	182351.6	853.7072
MissForest	0.007586413	0.01496116	185668.9	868.0717
XGBoost	0.008619215	0.01507733	202512.4	860.0427

Рис. 14: Результат

ВИСНОВКИ

В роботі було розглянуто і застосовано п’ять різних підходів до вирішення проблеми пропущених даних, а також обчислено нормалізовану середньоквадратичну похибку NRMSE та середню абсолютну похибку MAE оцінки ефективності методів.

Очевидно, що найкращу якість імпутації ми отримали за допомогою методу імпутації, що працює на основі алгоритму випадкового лісу missForest, найгірші показники отримано за допомогою методу швидкої заміни Hot deck. Такі результати не є випадковістю і подібні значення NRMSE та MAE ми отримаємо виконуючи імпутації для різних розподілів пропусків (тобто беручи різні значення seed при генерації номерів записів з яких вилучаються дані).

Причиною того, що метод швидкої заміни не є ефективним для даного масиву при імпутації неперервних змінних є природа даних, які мають широкий розмах вибірки і досить невелику кількістю спостережень в базі. Hot deck з самого початку був розроблений для заміни пропущених значень при проведенні анкетувань, тобто для роботи з даними більш зручними для категоризації за принципом, який цей алгоритм використовує.

Дещо кращі результати ми отримати за допомогою методу заміни середнім та kNN алгоритму, проте найбільш підходящими для даного випадку є методи машинного навчання, що використовують дерева прийняття рішень, при чому метод missForest, в основі якого лежить алгоритм Random Forest, дозволяє отримати прогнозовані значення максимально близькі до істинних, з чого можемо зробити висновок, що беггінг є більш вдалим способом прогнозування великих збитків при рідкісних подіях ніж бустинг, що лежить в основі методу максимального градієнтного підсилення.

БІБЛІОГРАФІЯ

[1] Little R. J. A., Rubin D. B. (2020), *Statistical Analysis with Missing Data,* Wiley, New Jersey

[2] Morgan L. (2020), *MissForest - missing data imputation using iterated random forests*

[3] König, T., Finke, D., & Daimer, S. (2005). *Ignoring the Non-ignorables? Missingness and Missing Positions*. European Union Politics, 6(3), 269-290.

[4] Maagorzata M. *IMPUTATION OF MISSING DATA USING R PACKAGE*. FOLIA OECONOMICA 269, 2012

[5] A. Kowarik, M. Templ (2016) Imputation with R package VIM. Journal of Statistical Software

[6] Andridge, R. R. and Little, R. J.A. (2010). *A Review of Hot Deck Imputation for Survey. Non-response. International Statistical Review*, 78(1), 40-64

[7] Laura B., Yarissa G., Katherine J. T. *Hot Deck Imputation: How Many Times Can or Should a Donor Be Used?* Int Stat Rev. 2010 Apr, 78(1): 40–64

[8] Dieter William Joenssen, Udo Bankhofer. *Hot Deck Methods for Imputing Missing Data The Effects of Limiting Donor Usage*. Technische Universität Ilmenau, Fachgebiet für Quantitative Methoden, Ilmenau, Germany

[9] Daniel J. Stekhoven, Peter Bühlmann. *MissForest—non-parametric missing value imputation for mixed-type data*. Bioinformatics, Volume 28, Issue 1, January 2012, Pages 112–118

[10] Richard Olshen. *A Conversation with Leo Breiman. Statistical Science.* 2001, Vol. 16, No. 2, 184–198

[11] Leo Breiman. *Random Forests*. Machine Learning, Volume 45, pages 5–32, 2001

[12] BIANCHI Blandine. *Application of the MissForest algorithm for imputation in the Survey on Income and Living Conditions.* UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS, 3-7 October 2022

[13] Adele Cutler, D. Richard Cutler and John R. Stevens. *Ensemble Machine Learning: Methods and Applications*. Machine Learning 45(1):157-176

[14] Yongshi Deng, Thomas Lumley. *Multiple Imputation Through XGBoost.* Journal of Computational and Graphical Statistics, Volume 33, 19 Oct 2023

[15] Chen, T., and Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System.* in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: Association for Computing Machinery, pp. 785–794