Контроль набутих знань
практика 5

Завдання

Для роботи використовуються файли Titanic.csv, train.csv, test.csv , titanic3. csv

Список обов'язкових кроків для виконання:

1. Підготовка даних

1.1. Створіть код в SAS Studio з назвою у форматі - lastname_cw5.sas;

1.2. Створіть у SAS Studio макрозмінні для свого SAS ID та прізвища.

1.3. Створіть бібліотеку(назва бібліотеки=прізвище інтерна), в якій будете зберігати лише фіналізовані датасети *(проміжні датасети зберігайте в директорії WORK).*

2. Формування цільового датасету

2.1. Імпортуйте всі необхідні дані у створену бібліотеку.

2.2. Створіть єдиний датасет Titanic_full, в якому будуть такі змінні:

Основні змінні з Titanic.csv

Додаткові змінні:PassengerId, sibsp, parch (із train/test); embarked, boat, body, home_dest (із titanic3.csv)

3. Створення нових ознак

3.1. Створіть нові змінні для подальшого аналізу:

FamilySize — розмір родини (формула: sibsp + parch + 1)

IsAlone — бінарна змінна (1 — пасажир подорожує один, 0 — не один)

Title — виділіть з імені соціальний титул (Mr, Mrs, Miss, Master, Dr, тощо)

Deck — літера палуби (перша літера з поля Cabin, якщо відсутнє — пропуск)

4. Попередня обробка даних

4.1. Заповніть пропущені значення:

Age, Fare — замініть пропуски на медіану

Embarked — замініть пропуски на найпопулярніше значення

4.2. Категоризуйте змінну Age у групи (змінна AgeGroup):

Child (0–12 років), Teen (13–18 років), Adult (19+ років)

4.3 *Висновок: Опишіть у вигляді коментарів як змінилася структура даних після заповнення пропусків і категоризації. Як розподілилися вікові групи?*

5. Дослідницький аналіз

5.1. Виведіть описову статистику для основних змінних (Age, Fare, FamilySize).

5.2. Побудуйте частотні таблиці для змінних (Gender, Class, Embarked, Title, Deck, IsAlone, AgeGroup, FamilySize).

5.3 *Висновок: Опишіть у вигляді коментарів які категорії переважають. Наприклад, яка стать або клас найчастіше зустрічається?*

6. Моделювання

6.1. Побудуйте модель логістичної регресії (PROC LOGISTIC) для прогнозування ймовірності виживання (Survived).

6.2. Включіть у модель такі змінні (усі або частину):

Age, Gender, Class, Embarked, AgeGroup, Fare, FamilySize, IsAlone, Title, Deck.

6.3. *Перевірте значущість змінних і визначте, які з них найбільше впливають на виживання – подати як коментар.*

7. Візуалізація

7.1. Побудуйте графіки розподілу виживаності по вибраних змінних (наприклад, по вікових групах, по розміру родини, по статі, по класу, по порту посадки).

7.2 *Висновок: Опишіть у вигляді коментарів що видно з графіків? Наприклад: "Жінки мали вищу виживаність", "Виживаність зростає з вищим класом", тощо.*

8. Дослідження взаємозв'язків між змінними

Візуалізуйте залежність між Fare та FamilySize для пасажирів, які вижили та не вижили (наприклад, розсіювання).

Висновок : Опишіть у вигляді коментарів який характер має залежність між Fare та FamilySize? Чи є кластери вижившиx з певними значеннями?

9. Моделювання та валідація

9.1. Проведіть розбиття датасету на тренувальну та тестову вибірки (наприклад, 70/30).

9.2 Побудуйте логістичну модель на тренувальній вибірці, а на тестовій — оцініть точність (accuracy), повноту (recall) та точність передбачення (precision).

9.3 Висновок : Опишіть у вигляді коментарів результати:

Accuracy: XX%, Recall: XX%, Precision: XX%, AUC: XX

"Модель на нових даних показала точність XX%. Це свідчить про …"

10. Висновки

10.1. Зробіть короткий висновок (1-2 абзаци): які фактори впливають на виживаність пасажирів, чи логічні ці результати, чи допомогли нові змінні покращити модель.

***Примітка**: Усі похідні числові змінні (наприклад, імовірності виживання, оцінки моделі, т.д.) мають бути округлені до двох десяткових знаків для забезпечення узгодженості та зручності у звітах.*