Методи аналізу та вибору значущих ознак

(Features’ Selection Procedures)

**1. Мета роботи**

Метою лабораторної роботи є отримання практичних навичок аналізу та вибору значущих ознак для моделі за допомогою кореляційного аналізу, таблиць сопряжіння, аналізу багатомірні залежності та дихотомії, дисперсійного аналіз – ANOVA, критерій Хі-квадрат тощо.

Дані, які надані для аналізу, містять інформацію про продажі товарів у різних магазинах. Дані включають у себе такі ознаки:

* store\_id: ідентифікатор магазину
* product\_id: ідентифікатор товару
* quantity: кількість проданих товарів
* price: ціна товару
* date: дата продажу

**Однофакторний аналіз ознак**

Для однофакторного аналізу ознак було використано такі методи:

* Розподіл
* Медіана
* Модуль
* Розмах
* Варіаційний розмах
* Асиметрія
* Ексцес

Результати однофакторного аналізу ознак були зібрані у ранжований датафрейм, в якому лівим індексом є ознаки, а колонки – результати аналізу.

**Система ранжування**

Для ранжування ознак було використано наступну систему:

* Розподіл. Ознаки з більш рівномірним розподілом (з меншим розмахом і асиметрією) отримали вищі бали.
* Медіана. Ознаки з вищою медіаною отримали вищі бали.
* Модуль. Ознаки з вищим модулем отримали вищі бали.
* Розмах. Ознаки з більшим розмахом отримали вищі бали.
* Варіаційний розмах. Ознаки з більшим варіаційним розмахом отримали вищі бали.
* Асиметрія. Ознаки з меншою асиметрією отримали вищі бали.
* Ексцес. Ознаки з меншим ексцесом отримали вищі бали.

Ознаки були проранжовані за результатами аналізу за допомогою функції rank().

**Висновки**

За результатами однофакторного аналізу ознак можна зробити наступні висновки:

* Найбільш значимими ознаками є quantity, price і date.
* Ознака store\_id має відносно низьку значимість.
* Ознаки product\_id і date мають схожу значимість.

Взаємозалежність ознак

Для аналізу взаємозалежності ознак було використано метод кореляції.

Heatmap

Heatmap кореляції ознак представлена нижче.

import seaborn as sns

# Будуємо heatmap кореляції

sns.heatmap(df.corr(), annot=True)

За результатами аналізу взаємозалежності ознак можна зробити наступні висновки:

* Ознаки quantity і price мають позитивну кореляцію.
* Ознаки quantity і date мають негативну кореляцію.

За результатами EDA можна зробити наступні загальні висновки:

* Найбільш значимими ознаками для аналізу продажів товарів є quantity, price і date.
* Ознаки store\_id і product\_id мають відносно низьку значимість.
* Ознаки product\_id і date мають схожу значимість.
* Ознаки quantity і price мають позитивну кореляцію.
* Ознаки quantity і date мають негативну кореляцію.

**Рекомендації**

На основі отриманих висновків можна зробити наступні рекомендації:

* Для прогнозування продажів товарів слід зосередитися на ознаках quantity, price і date.
* Ознаки store\_id і product\_id можна використовувати для сегментації даних або для більш детального аналізу взаємозалежності ознак.