ИУ5-65Б Вопияшин Никита РК2 ТМО, Вариант 3

Задание:
Для заданного набора данных (по Вашему варианту) постройте модели классификации или регрессии (в зависимости от конкретной задачи, рассматриваемой в наборе данных). Для построения моделей используйте методы 1 и 2 (по варианту для Вашей группы). Оцените качество моделей на основе подходящих метрик качества (не менее двух метрик). Какие метрики качества Вы использовали и почему? Какие выводы Вы можете сделать о качестве построенных моделей? Для построения моделей необходимо выполнить требуемую предобработку данных: заполнение пропусков, кодирование категориальных признаков, и т.д.

Методы:
метод опорных векторов, градиентный бустинг

Предварительная обработка данных:

```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from sklearn.svm import SVR
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import mean_squared_error, mean_absolute_error, r2_score
from sklearn.impute import SimpleImputer

data = pd.read_csv('HousingData.csv')

print(data.isnull().sum())

imputer = SimpleImputer(strategy='median')
data_imputed = pd.DataFrame(imputer.fit_transform(data), columns=data.columns)

features = ['CRIM', 'ZN', 'INDUS', 'NOX', 'RM', 'AGE', 'DIS', 'TAX', 'PTRATIO', 'LSTAT']
X = data_imputed[features]
y = data_imputed['MEDV']

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Метод опорных векторов:

```
svr = SVR(kernel='rbf', C=1.0, epsilon=0.1)
svr.fit(X_train_scaled, y_train)

y_pred_svr = svr.predict(X_test_scaled)

mse_svr = mean_squared_error(y_test, y_pred_svr)
mae_svr = mean_absolute_error(y_test, y_pred_svr)
r2_svr = r2_score(y_test, y_pred_svr)

print(f"SVR - MSE: {mse_svr:.2f}, MAE: {mae_svr:.2f}, R2: {r2_svr:.2f}")
```

```
SVR - MSE: 25.75, MAE: 2.77, R2: 0.65
```

Градиентный бустинг:

```
gbr = GradientBoostingRegressor(n_estimators=100, learning_rate=0.1, max_depth=3, random_state=42)
gbr.fit(X_train, y_train)

y_pred_gbr = gbr.predict(X_test)

mse_gbr = mean_squared_error(y_test, y_pred_gbr)
mae_gbr = mean_absolute_error(y_test, y_pred_gbr)
r2_gbr = r2_score(y_test, y_pred_gbr)

print(f"Gradient Boosting - MSE: {mse_gbr:.2f}, MAE: {mae_gbr:.2f}, R2: {r2_gbr:.2f}")
```

```
Gradient Boosting - MSE: 7.50, MAE: 1.93, R2: 0.90
```

Использованные метрики

1. Mean Squared Error (MSE) - средняя квадратичная ошибка:
    - Штрафует большие ошибки сильнее, чем маленькие
    - Чем меньше значение, тем лучше модель
    - Единицы измерения: квадрат единиц целевой переменной (в данном случае, $1000²)
2. Mean Absolute Error (MAE) - средняя абсолютная ошибка:

- Показывает среднюю величину ошибки в тех же единицах, что и целевая переменная
- Более интерпретируема, чем MSE
- Единицы измерения: $1000

3. R^2 (коэффициент детерминации):
   - Показывает, какая доля дисперсии зависимой переменной объясняется моделью
   - Диапазон значений: от 0 до 1 (чем ближе к 1, тем лучше)
   - Безразмерная величина

Почему выбраны именно такие метрики:

- MSE выбрана как стандарт для сравнения алгоритмов (особенно SVR, где оптимизируется квадратичная функция потерь).
- MAE добавлена для "человеко-читаемой" интерпретации ошибок в долларах.
- R^2 помогает понять, насколько хорошо признаки объясняют целевую переменную, что важно для feature engineering.

Интерпретация результатов:

1. Метод опорных векторов (SVR):
   - MSE = 25.75 означает, что средний квадрат ошибки составляет 25.75 (в единицах $1000²)
   - MAE = 2.77 означает, что средняя ошибка предсказания составляет $2,770
   - R^2 = 0.65 означает, что модель объясняет 65% вариативности данных

1. Градиентный бустинг
    - MSE = 7.50 (значительно лучше, чем у SVR)
    - MAE = 1.93 (ошибка около $1,930, что на 30% лучше, чем SVR)
    - R^2 = 0.90 (отличный результат, модель объясняет 90% вариативности)

Выводы:

Градиентный бустинг показал значительно лучшие результаты по всем метрикам:

- MSE в 3.4 раза меньше, чем у SVR
- MAE на 30% лучше
- R^2 на 25 процентных пунктов выше

Причины различий в качестве моделей

1. Градиентный бустинг:
    - Лучше справляется с нелинейными зависимостями
    - Автоматически учитывает взаимодействия признаков
    - Более устойчив к шуму в данных
1. Метод опорных векторов:
    - Может требовать более тщательной настройки гиперпараметров
    - Чувствителен к масштабированию данных
    - Менее гибок в моделировании сложных зависимостей