Ходырев Роман Владиславович

ИУ5-65Б

18 вариант

In [1]:
```
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.svm import SVR
from sklearn.ensemble import GradientBoostingRegressor
from sklearn.metrics import mean_absolute_error, mean_squared_error
from sklearn.preprocessing import LabelEncoder
```

In [3]:
```
df = pd.read_csv('investments_VC_regression.csv', encoding='latin1', sep=None, engine='python')
```

In [4]:
```
df.columns = df.columns.str.strip()
```

In [5]:
```
df = df[['funding_total_usd', 'country_code', 'funding_rounds', 'founded_year']]
```

In [6]:
```
df['funding_total_usd'] = df['funding_total_usd'].replace('[\$,]', '', regex=True).replace('None', np.nan)
df['funding_total_usd'] = pd.to_numeric(df['funding_total_usd'], errors='coerce')
df['funding_total_usd'] = df['funding_total_usd'].fillna(df['funding_total_usd'].median())
```

In [7]:
```
df['founded_year'] = pd.to_numeric(df['founded_year'], errors='coerce')
df['founded_year'] = df['founded_year'].fillna(df['founded_year'].median())
df['funding_rounds'] = df['funding_rounds'].fillna(df['funding_rounds'].median())
df['country_code'] = df['country_code'].fillna('UNKNOWN')
```

In [8]:
```
le = LabelEncoder()
df['country_code'] = le.fit_transform(df['country_code'])
```

In [9]:
```
X = df.drop('funding_total_usd', axis=1)
y = df['funding_total_usd']
```

In [10]:
```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

In [11]:
```
svr = SVR()
svr.fit(X_train, y_train)
y_pred_svr = svr.predict(X_test)
```

In [12]:
```
gbr = GradientBoostingRegressor()
gbr.fit(X_train, y_train)
y_pred_gbr = gbr.predict(X_test)
```

In [13]:
```
def print_metrics(y_true, y_pred, model_name):
    mae = mean_absolute_error(y_true, y_pred)
    rmse = np.sqrt(mean_squared_error(y_true, y_pred))
    print(f'{model_name}:\nMAE: {mae:.2f}\nRMSE: {rmse:.2f}\n')
```

In [14]:
```
print_metrics(y_test, y_pred_svr, 'SVR')
print_metrics(y_test, y_pred_gbr, 'Gradient Boosting Regressor')
```

```
SVR:
MAE: 10405733.69
RMSE: 42719576.40

Gradient Boosting Regressor:
MAE: 12857368.24
RMSE: 41356065.70
```

Классификация или регессия?

В данном случае используется регрессия. Мы предсказывали не категориальную переменную, а непрерывную величину — размер финансирования (funding_total_usd), что и определяет тип задачи как регрессионный.

Какие метрики качества Вы использовали и почему?

Мы использовали две основные метрики для оценки качества регрессионных моделей:

- MAE (Mean Absolute Error) — средняя абсолютная ошибка:
  - Показывает среднюю величину отклонения предсказания от фактического значения.

    - Удобна тем, что измеряется в тех же единицах, что и целевая переменная.
    - Устойчива к выбросам, в отличие от RMSE.
- RMSE (Root Mean Squared Error) — среднеквадратичная ошибка:
    - Более чувствительна к крупным ошибкам, потому что ошибки возводятся в квадрат.
    - Помогает понять, насколько сильно модель может ошибаться в наихудших случаях.
    - Хорошо показывает наличие/влияние выбросов.

Выбор этих двух метрик позволяет объективно оценить качество модели: MAE показывает среднюю точность, RMSE — чувствительность к ошибкам.

Какие выводы можно сделать о качестве построенных моделей?

- Обе модели (SVR и Gradient Boosting Regressor) показали сравнимые результаты по MAE и RMSE:
    - SVR:
        - MAE ≈ 10.4 млн,
        - RMSE ≈ 42.7 млн
    - Градиентный бустинг:
        - MAE ≈ 12.8 млн,
        - RMSE ≈ 41.3 млн
- SVR показал более низкое среднее абсолютное отклонение (MAE), но более высокое RMSE, что говорит о том, что он чаще делает более точные предсказания, но может сильно ошибаться на некоторых выбросах.
- Gradient Boosting оказался немного устойчивее к выбросам, судя по чуть более низкому RMSE, но в среднем ошибался сильнее (выше MAE).

Вывод:

- Оба метода справились средне — ошибки довольно большие (десятки миллионов долларов).
- Вероятно, распределение целевой переменной (funding_total_usd) имеет много выбросов, что делает задачу сложной.
- Можно попробовать улучшить модели путём:
    - логарифмирования целевой переменной,
    - отбора фичей,
    - нормализации данных,
    - работы с выбросами.