Regression Model Report

Kevin Olind Hasudungan Nainggolan/1103210140

Exploratory Data Analysis (EDA)

- **Dataset Summary**:
    - Dataset 515,345 baris dan 91 kolom
    - Dikarenakan tidak adanya header jadi saat memanggil data dari drive digunakan **header=None** untuk memberitahu code bahwa tidak ada header jadi baris pertama dari dataset tidak disalah artikan menjadi header
    - Lalu untuk memperbagus data diberikan nama x2,x3,x4,x5,… dan karena ada satu kolom yang terlihat seperti data tahun maka kolom tersebut diberikan nama Year
    - Setelah dilakukan scan data dengan nilai NaN/NULL dengan ***df.isna().sum()*** tidak didapati nilai kosong didalam data, jadi data terisi semua dengan nilai/value.
    - Dataset ini juga terdiri dari data numerik tipe float64 dan int64 untuk kolom Year
- **Distribusi Variable Tahun**:
    - Dilakukan pengecekan *instances* pada kolom Year dengan ***df['Year'].value_counts()***.
    - Terdapat 10 tahun yang paling sering keluar dari tahun 2000 sampai 2007.
    - Tahun 2007 sering keluar didalam data sebanyak 39000 kali, sedangkan tahun 2011 hanya memiliki 1 *instance*.
- **Korelasi Antar Kolom (Heatmap)**:
    - Dikarenakan banyaknya data yang ada maka saya menambahkan syarat nilai yang ditampilkan hanya nilai dengan nilai > 0.5 dan nilai < -0.2.
    - Meskipun sudah difilter tetapi karena banyaknya data dan hanya sedikit yang memiliki korelasi jadi hasil yang didapat kurang terlihat
    - Setelah saya lihat satu persatu saya mencoba menggunakan nilai dengan korelasi yang memiliki nilai korelasi seperti persyarat yaitu ['x3', 'x9', 'x15', 'x16', 'x17', 'x19', 'x20', 'x21', 'x22', 'x23', 'x24', 'x25', 'x47', 'x57', 'x61', 'x64']

Hasil Hyperparameter Tuning

a. Polynomial Regression

Proses ini menggunakan 3-fold cross-validation, sehingga setiap kombinasi parameter diuji sebanyak tiga kali pada subset data yang berbeda. Total pengujian yang dilakukan adalah 12 fits (4 kandidat parameter × 3 fold). Setiap fold bertindak sebagai validasi, sedangkan sisanya menjadi training set.

- Hyperparameter yang Dilakukan tuning: Derajat polinomial (poly_features__degree).
- Parameter Terbaik: degree = 2
- Skor R-squared Terbaik: 0.0607

Meskipun skor R^2 tersebut rendah. Hal ini menunjukkan bahwa model hanya mampu menjelaskan sekitar 6% dari variasi dalam data target (Year).

b. Decision Tree Regressor

Dalam proses tuning ini, terdapat dua hyperparameter yang diuji, yaitu max_depth dan min_samples_split. Kombinasi dari kedua parameter ini menghasilkan total 9 kandidat parameter yang diuji. Dengan menggunakan 5-fold cross-validation, setiap kombinasi diuji pada 5 subset data yang berbeda, sehingga total terdapat 45 fits (9 kandidat × 5 fold).

- Hyperparameter yang Dilakukan tuning:
  - max_depth: [3, 5, 10]
  - min_samples_split: [2, 5, 10]
- Parameter Terbaik:
  - max_depth=10
  - min_samples_split=10
- Skor R-squared Terbaik: 0.0689

Skor R^2 ini menunjukkan bahwa model dapat menjelaskan sekitar 6.89% variasi dalam data target (Year), yang merupakan peningkatan kecil dibandingkan model lain seperti Polynomial Regression.

c. k-Nearest Neighbors (k-NN)

- Hyperparameter yang Dituning:
  - n_neighbors: [3, 5, 7]
  - weights: ['uniform', 'distance']
  - model_p: [1, 2]
- Parameter Terbaik:
  - n_neighbors: 7
  - p: 2 (Euclidean Distance)
  - weights: 'distance'
- Skor R-squared Terbaik: 0.0689

  Ini menunjukkan bahwa model hanya mampu menjelaskan sekitar 0.93% variasi dalam target data (Year), yang sangat kecil. Hal ini mengindikasikan bahwa hubungan antara fitur dan target tidak cukup kuat untuk ditangkap oleh k-NN.

d. XGBoost

- Hyperparameter yang Dituning:
  - n_estimators: [50, 100, 200]
  - max_depth: [3, 5, 7]
  - learning_rate: [0.01, 0.1, 0.2]
  - subsample: [0.8, 1.0]
  - colsample_bytree: [0.8, 1.0]
- **Parameter Terbaik**:
  - colsample_bytree: 1.0 (menggunakan semua fitur untuk membangun setiap pohon)
  - learning_rate: 0.1 (learning rate yang cukup konservatif)
  - max_depth: 5 (kedalaman maksimum pohon)
  - n_estimators: 200 (jumlah pohon yang digunakan)
  - subsample: 0.8 (menggunakan 80% data untuk setiap pohon)
- **Skor R^2 Terbaik**: **0.1155**.

Kombinasi ini dipilih oleh GridSearchCV karena memberikan performa terbaik berdasarkan validasi silang. Ini adalah skor R-squared terbaik sejauh ini dibandingkan dengan model lain seperti Polynomial Regression, Decision Tree, dan k-NN. Skor ini menunjukkan bahwa XGBoost mampu menjelaskan sekitar 11.55% variasi dalam data target (Year).

Kesimpulan

Dari hasil hyperparameter tuning pada empat model regresi yang diuji (Polynomial Regression, Decision Tree, k-NN, dan XGBoost), terlihat bahwa XGBoost Regressor memberikan performa terbaik dengan nilai R^2 sebesar 0.1155, dibandingkan model lainnya. Decision Tree berada di urutan kedua dengan R^2 sebesar 0.0689, diikuti oleh Polynomial Regression 0.0608 dan k-NN 0.0093. XGBoost unggul karena kemampuannya menangkap hubungan non-linear yang kompleks melalui pendekatan ensemble learning, sementara Decision Tree menunjukkan performa yang baik dalam menangkap pola sederhana namun cenderung overfit. Di sisi lain, k-NN menunjukkan performa terendah karena sifatnya yang bergantung pada hubungan lokal antar titik data, yang tampaknya tidak sesuai dengan karakteristik dataset ini.