Kinanti Aria Widaswara
1103213165

1. Jika model linear regression atau decision tree mengalami underfitting pada dataset ini, strategi apa yang akan digunakan untuk meningkatkan performanya? Bandingkan setidaknya dua pendekatan berbeda (misal: transformasi fitur, penambahan features, atau perubahan model ke algoritma yang lebih kompleks), dan jelaskan bagaimana setiap solusi memengaruhi bias-variance tradeoff!

a. Transformasi Fitur:

- **Polynomial Features:** Menambahkan fitur polinomial (misalnya, kuadrat atau kubik dari fitur asli) dapat membantu model linear menangkap hubungan non-linear dalam data. Namun, ini dapat meningkatkan kompleksitas model dan risiko overfitting jika tidak dikendalikan dengan baik.
- **Transformasi Logaritmik atau Eksponensial:** Menerapkan transformasi seperti logaritma atau eksponensial pada fitur atau target dapat membantu mengatasi non-linearitas dan heteroskedastisitas dalam data.

b. Penambahan Fitur:

- **Interaksi Antar Fitur:** Menambahkan fitur yang merupakan hasil interaksi antara dua atau lebih fitur asli dapat membantu model menangkap hubungan kompleks dalam data.
- **Eksternal Data:** Mengintegrasikan data eksternal yang relevan dapat memperkaya informasi dan membantu model memahami pola yang lebih kompleks.

c. Penggunaan Model yang Lebih Kompleks:

- **Model Ensemble:** Menggunakan model seperti Random Forest atau Gradient Boosting dapat menangkap hubungan non-linear dan interaksi kompleks antar fitur.
- **Support Vector Regression (SVR):** SVR dengan kernel non-linear dapat menangkap pola yang tidak dapat ditangkap oleh model linear sederhana.

Pengaruh terhadap Bias-Variance Tradeoff:

- **Transformasi Fitur dan Penambahan Fitur:** Dapat mengurangi bias dengan menangkap pola yang lebih kompleks, namun dapat meningkatkan varians jika model menjadi terlalu kompleks.
- **Model yang Lebih Kompleks:** Cenderung memiliki varians yang lebih tinggi, sehingga penting untuk menggunakan teknik regularisasi dan validasi silang untuk menghindari overfitting.

2. Selain MSE, jelaskan dua alternatif loss function untuk masalah regresi (misal: MAE, Huber loss) dan bandingkan keunggulan serta kelemahannya. Dalam skenario apa setiap loss function lebih cocok digunakan? (Contoh: data dengan outlier, distribusi target non-Gaussian, atau kebutuhan interpretasi model).

a. Mean Absolute Error (MAE):

- **Keunggulan:** Lebih robust terhadap outlier dibandingkan MSE karena tidak mengkuadratkan error.
- **Kelemahan:** Tidak differentiable di titik nol, yang dapat menyulitkan optimisasi menggunakan metode berbasis gradien.
- **Cocok untuk:** Data dengan outlier signifikan atau ketika penalti besar terhadap outlier tidak diinginkan.

b. Huber Loss:

- **Keunggulan:** Kombinasi antara MSE dan MAE; berperilaku seperti MSE untuk error kecil dan seperti MAE untuk error besar, sehingga lebih robust terhadap outlier.
- **Kelemahan:** Memerlukan penentuan parameter delta, yang menentukan titik transisi antara perilaku MSE dan MAE.
- **Cocok untuk:** Situasi di mana terdapat outlier, tetapi model tetap perlu sensitif terhadap error kecil.

3. Tanpa mengetahui nama fitur, metode apa yang dapat digunakan untuk mengukur pentingnya setiap fitur dalam model? Jelaskan prinsip teknikal di balik metode tersebut (misal: koefisien regresi, feature importance berdasarkan impurity reduction) serta keterbatasannya!

a. Koefisien Regresi:

- **Prinsip:** Dalam regresi linear, koefisien menunjukkan pengaruh setiap fitur terhadap target. Koefisien yang lebih besar (dalam nilai absolut) menunjukkan fitur yang lebih penting.
- **Keterbatasan:** Hanya berlaku untuk model linear dan dapat dipengaruhi oleh multikolinearitas antar fitur.

b. Feature Importance Berdasarkan Impurity Reduction:

- **Prinsip:** Dalam model pohon keputusan, fitur yang sering digunakan untuk membagi data dan menghasilkan pengurangan impurity yang besar dianggap lebih penting.
- **Keterbatasan:** Dapat bias terhadap fitur dengan banyak kategori atau nilai unik .

c. Permutation Importance:

- **Prinsip:** Mengukur penurunan kinerja model saat nilai fitur diacak. Penurunan kinerja yang signifikan menunjukkan fitur yang penting.
- **Keterbatasan:** Memerlukan komputasi tambahan dan dapat dipengaruhi oleh korelasi antar fitur.

4. Bagaimana mendesain eksperimen untuk memilih hyperparameter optimal (misal: learning rate untuk SGDRegressor, max_depth untuk Decision Tree) pada dataset ini? Sertakan analisis tradeoff antara komputasi, stabilitas pelatihan, dan generalisasi model!

a. Teknik Tuning:

- **Grid Search:** Mencoba semua kombinasi hyperparameter dalam rentang tertentu. Akurat tetapi komputasi intensif.
- **Random Search:** Memilih kombinasi hyperparameter secara acak. Lebih efisien secara komputasi dibandingkan grid search.
- **Bayesian Optimization:** Menggunakan model probabilistik untuk memprediksi kombinasi hyperparameter terbaik berdasarkan hasil sebelumnya.

b. Pertimbangan Trade-off:

- **Komputasi:** Grid search dapat menjadi sangat mahal secara komputasi, terutama dengan banyak hyperparameter.
- **Stabilitas Pelatihan:** Hyperparameter seperti learning rate yang terlalu tinggi dapat menyebabkan pelatihan tidak stabil.
- **Generalisasi Model:** Hyperparameter harus dipilih untuk memaksimalkan kinerja pada data yang tidak terlihat, bukan hanya data pelatihan.

5. Jika menggunakan model linear regression dan residual plot menunjukkan pola non-linear serta heteroskedastisitas, langkah-langkah apa yang akan diambil? (contohnya: Transformasi data/ubah model yang akan dipakai/etc)

a. Transformasi Data:

- **Transformasi Logaritmik atau Box-Cox:** Dapat membantu menstabilkan varians dan membuat hubungan lebih linear.
- **Transformasi Yeo-Johnson:** Alternatif dari Box-Cox yang dapat menangani nilai nol dan negatif.

b. Ubah Model:

- **Polynomial Regression:** Menambahkan fitur polinomial untuk menangkap hubungan non-linear.

- **Model Non-linear:** Menggunakan model seperti Random Forest atau SVR yang dapat menangkap pola non-linear tanpa perlu transformasi fitur.

c. Weighted Least Squares (WLS):

- **Prinsip:** Memberikan bobot lebih kecil pada observasi dengan varians yang lebih tinggi, membantu mengatasi heteroskedastisitas.

d. Heteroskedasticity-Consistent Standard Errors (HCSE):

- **Prinsip:** Menyesuaikan standar error untuk mengakomodasi heteroskedastisitas tanpa mengubah koefisien model.