Analisa Regresi Model

Nama: Arif Al Imran

Kelas: TK-45-G05

NIM: 1103210193

Soal:

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 (misal: transformasi fitur, penambahan fitur, atau perubahan model ke algoritma yang lebih kompleks), dan jelaskan bagaimana setiap Solusi mengaruhi bias-vatiance tradeoff!
2. Selain MSE, jelaskan dua alternatif loss function untuk masalah regresi (misal: MAE, Huber loss) dan bandingkan keunggulan serta kelemahannya. Dalam scenario apa setiap loss function lebih cocok digunakan? (Contoh: data dengan outlier, distribusi target non-Gaussian, atau kebutuhan interpretasi model).
3. Tanpa mengetahui nama fitur, metode apa yang dpat 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!
4. Bagaimana mendesain eksperimen untuk memilih hyperparameter optimal (misal: learning rate untuk SGDRegressor, max_depth untuk Decision Tree) pada dataset ini? Sertakan analisis trade off antara komputasi, stabilitas pelatihan, dan generalisasi model!
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 diapakai)

Jawab:

1. Strategi Mengatasi Underfitting

Untuk mengatasi underfitting pada model linear regression atau decision tree, dua strategi utama dapat diterapkan:

Rekayasa Fitur:

- Menambahkan fitur polinomial (x^2, x^3) atau interaksi antar fitur
- Menerapkan transformasi non-linear pada fitur
- Keuntungan: Mempertahankan interpretabilitas model sambil mengurangi bias
- Risiko: Meningkatkan variance jika tidak diregularisasi dengan tepat

Peningkatan Kompleksitas Model:

- Linear regression: Beralih ke regresi polinomial, SVR dengan kernel non-linear, atau metode ensemble
- Decision tree: Memperbesar max_depth, mengurangi min_samples_split, atau menurunkan threshold impurity
- Keuntungan: Langsung mengurangi bias untuk menangkap pola kompleks
- Risiko: Potensi overfitting jika tidak divalidasi dengan baik

Dalam perspektif bias-variance tradeoff, rekayasa fitur umumnya lebih baik untuk mempertahankan interpretabilitas. Namun, untuk hubungan yang sangat kompleks, peningkatan kompleksitas model mungkin diperlukan meskipun interpretabilitas berkurang.

2. Alternatif Loss Function untuk Regresi

Mean Absolute Error (MAE):

- Dihitung sebagai rata-rata nilai absolut dari selisih prediksi dan nilai aktual
- Keunggulan: Lebih tahan terhadap outlier, mengoptimalkan nilai median, interpretasi mudah (unit sama dengan variabel target)
- Kelemahan: Tidak diferensiabel di titik nol, dapat menghambat optimasi berbasis gradien
- Cocok untuk: Data dengan outlier signifikan atau distribusi error non-Gaussian

Huber Loss:

- Fungsi hybrid: Seperti MSE untuk error kecil ($|y-\hat{y}| \leq \delta$) dan seperti MAE untuk error besar
- Keunggulan: Menyeimbangkan diferensiabilitas MSE dengan ketahanan MAE terhadap outlier
- Kelemahan: Memerlukan penyetelan parameter δ tambahan dan komputasi lebih kompleks
- Cocok untuk: Dataset dengan beberapa outlier atau saat membutuhkan optimasi berbasis gradien yang tahan outlier

3. Metode Pengukuran Pentingnya Fitur

Permutation Importance:

- Prinsip: Mengacak nilai setiap fitur secara bergantian dan mengukur dampaknya pada performa model
- Fitur yang pengacakannya menyebabkan penurunan performa terbesar dianggap paling penting
- Keunggulan: Universal, dapat diterapkan pada hampir semua model machine learning
- Keterbatasan: Komputasi intensif dan berpotensi meremehkan pentingnya fitur yang saling berkorelasi tinggi

Feature Importance dari Struktur Model:

- Linear regression: Menggunakan koefisien terstandarisasi sebagai ukuran langsung pentingnya fitur
- Decision tree: Mengukur berdasarkan pengurangan impurity saat fitur digunakan untuk splitting
- Keunggulan: Interpretasi langsung dan efisien secara komputasi
- Keterbatasan: Bergantung pada jenis model, bias terhadap fitur kardinalitas tinggi (tree), sensitif terhadap multikolinearitas (model linear)

4. Desain Eksperimen untuk Optimasi Hyperparameter

Optimasi hyperparameter dapat dilakukan melalui tiga pendekatan utama:

Grid Search:

- Evaluasi sistematis semua kombinasi dari nilai diskrit untuk setiap hyperparameter
- Keunggulan: Pemahaman komprehensif tentang landscape parameter
- Keterbatasan: Biaya komputasi tinggi, terutama untuk ruang parameter berdimensi tinggi

Randomized Search:

- Mengambil sampel acak dari distribusi parameter
- Keunggulan: Lebih efisien daripada Grid Search, mengalokasikan lebih banyak sumber daya untuk parameter berpengaruh
- Keterbatasan: Tidak menjamin menemukan kombinasi optimal

Bayesian Optimization:

- Menggunakan informasi dari evaluasi sebelumnya untuk membangun model hubungan antara hyperparameter dan performa
- Keunggulan: Paling efisien untuk menemukan parameter optimal
- Keterbatasan: Implementasi lebih kompleks

Analisis trade-off:

- Komputasi: Grid Search paling mahal, Bayesian paling efisien
- Stabilitas pelatihan: Cross-validation diperlukan untuk semua metode
- Generalisasi model: Validasi bersarang membantu memastikan hyperparameter yang dipilih menghasilkan model yang generalizeable

5. Mengatasi Non-Linearitas dan Heteroskedastisitas pada Regresi Linear

Ketika residual plot menunjukkan pola non-linear dan heteroskedastisitas, beberapa strategi efektif adalah:

Transformasi Variabel Target:

- Mengubah skala variabel dependen (log, akar kuadrat, Box-Cox)
- Keuntungan: Dapat melinearkan hubungan dan menstabilkan varians error
- Catatan: Memerlukan transformasi balik untuk interpretasi hasil

Transformasi Fitur:

- Memodifikasi variabel independen dengan transformasi logaritmik untuk hubungan eksponensial atau transformasi polinomial untuk hubungan melengkung
- Keuntungan: Mempertahankan skala asli variabel target
- Catatan: Memerlukan pemahaman domain yang baik

Model Linear yang Lebih Fleksibel:

- Generalized Linear Models dengan fungsi link yang sesuai
- Weighted Least Squares untuk mengatasi heteroskedastisitas
- Keuntungan: Mempertahankan interpretabilitas model linear

Beralih ke Model Non-Linear:

- Menggunakan Random Forest, Gradient Boosted Trees, atau SVR dengan kernel non-linear
- Keuntungan: Performa prediktif lebih baik untuk hubungan kompleks
- Trade-off: Interpretabilitas yang lebih rendah