Nama : Chandra Aulia Haswangga

NIM : 1103223163

Kelas : TK-45-G05

Tugas : UTS – Analisis Regresi

1. Strategi untuk Mengatasi Underfitting pada Linear Regression atau Decision Tree
    a) Transformasi Fitur
    Salah satu cara untuk menangani underfitting adalah dengan mengubah fitur yang ada melalui **transformasi non-linear**. Misalnya, jika data memiliki hubungan non-linear dengan target variabel, kita dapat menggunakan transformasi seperti **log**, **polynomial features**, atau **binerisasi** fitur.

    Pengaruh pada Bias-Variance Tradeoff:
    - Bias
    Transformasi fitur dapat menurunkan bias model dengan memperkenalkan lebih banyak informasi atau variabilitas dalam fitur yang memungkinkan model belajar pola yang lebih kompleks.
    - Variance
    Namun, jika transformasi terlalu banyak menambah fitur, hal ini dapat menyebabkan peningkatan variance dan berisiko terhadap overfitting jika jumlah data terbatas.

    b) Penambahan Fitur
    Penambahan fitur baru yang relevan (misalnya, fitur interaksi atau fitur eksternal) dapat membantu model memahami lebih banyak informasi tentang hubungan dalam data. Misalnya, menambahkan fitur yang menggabungkan dua atau lebih fitur yang ada dapat memberikan model lebih banyak informasi.

    Pengaruh pada Bias-Variance Tradeoff:
    - Bias
    Penambahan fitur yang relevan dapat mengurangi bias model, membuatnya lebih fleksibel dalam menangkap pola yang lebih rumit.

    - Variance
    Penambahan terlalu banyak fitur dapat menyebabkan overfitting, meningkatkan variance jika model terlalu rumit untuk jumlah data yang ada.

    c) Perubahan Model ke Algoritma yang Lebih Kompleks
    Jika model linear regression atau decision tree mengalami underfitting, kita bisa mencoba model yang lebih kompleks, seperti **Random Forest** atau **Gradient Boosting Machines (GBM)**. Kedua algoritma ini mengatasi keterbatasan decision tree dengan menggabungkan banyak pohon keputusan atau model.

Pengaruh pada Bias-Variance Tradeoff:

- Bias
  Model yang lebih kompleks biasanya dapat mengurangi bias karena mereka lebih fleksibel dalam menangkap pola non-linear dalam data.

- Variance
  Namun, algoritma yang lebih kompleks sering kali meningkatkan variance karena cenderung lebih sensitif terhadap fluktuasi data dan lebih rentan terhadap overfitting jika tidak diatur dengan benar.

2. Alternatif Loss Functions untuk Masalah Regresi
   a) Mean Absolute Error (MAE)
   MAE mengukur rata-rata absolut dari perbedaan antara nilai prediksi dan nilai aktual. Ini memberikan penalti yang sama untuk kesalahan besar maupun kecil.
      - **Keunggulan**:
        Robust terhadap outlier: MAE memberikan penalti yang lebih kecil terhadap outlier dibandingkan dengan MSE, sehingga lebih cocok ketika data memiliki banyak nilai ekstrem.
      - **Kelemahan**:
        MAE tidak memberi bobot lebih pada kesalahan besar, yang dapat menyebabkan model tidak memprioritaskan kesalahan besar yang lebih penting.
      - **Kapan digunakan**:
        MAE lebih cocok digunakan ketika data memiliki **outlier** atau distribusi target **non-Gaussian**.

   b) Huber Loss
   Huber loss menggabungkan MSE dan MAE. Ketika kesalahan kecil, ia bertindak seperti MSE (menghukum kesalahan kecil secara kuadrat), tetapi ketika kesalahan besar, ia berfungsi seperti MAE.
      - **Keunggulan**:
        Menangani outlier: Huber loss sangat berguna ketika dataset mengandung outlier, karena mengurangi pengaruh kesalahan besar sambil tetap memberi penalti untuk kesalahan kecil.
      - **Kelemahan**:
        Pilihan nilai parameter delta dapat memengaruhi kinerja dan interpretasi model.
      - **Kapan digunakan**:
        Huber loss lebih cocok digunakan ketika dataset mengandung **outlier** dan Anda ingin keseimbangan antara penalti kesalahan kecil dan besar.

3. Metode Mengukur Pentingnya Fitur dalam Model
   a) Koefisien Regresi (untuk Linear Regression)
      - Prinsip Teknis

Dalam regresi linear, koefisien fitur menunjukkan pentingnya masing-masing fitur terhadap target variabel. Fitur dengan koefisien besar memiliki pengaruh lebih besar terhadap prediksi model.

- Keterbatasan
  Koefisien hanya berlaku untuk model linear dan mungkin tidak memberikan informasi yang akurat untuk model non-linear.

b) Feature Importance Berdasarkan Impurity Reduction (untuk Decision Trees dan Random Forests)
   - Prinsip Teknis
     Dalam decision tree, fitur yang lebih sering digunakan untuk memecah data atau mengurangi impuritas (impurity) memiliki importance yang lebih tinggi. Random Forest menghitung importance rata-rata fitur dengan mengaggregasi hasil dari beberapa pohon.

   - Keterbatasan
     Feature importance tidak selalu dapat diinterpretasikan secara langsung dalam model non-tree seperti regresi linear. Selain itu, dalam beberapa kasus, fitur yang sangat berkorelasi dapat menyebabkan fitur tertentu dianggap lebih penting dari yang sebenarnya.

4. Eksperimen untuk Memilih Hyperparameter Optimal
   a) **Grid Search**
      Grid Search adalah metode yang mencoba berbagai kombinasi dari hyperparameter yang telah ditentukan untuk menemukan yang terbaik.
      - Trade-off:
        Memakan waktu karena menguji setiap kemungkinan, terutama jika dataset besar. Namun, ini memberikan hasil yang lebih sistematis dan dapat diandalkan.

   b) **Random Search**
      Random Search memilih kombinasi hyperparameter secara acak, yang lebih efisien dibandingkan Grid Search, terutama ketika ruang hyperparameter sangat besar.
      - Trade-off:
        Meskipun lebih cepat, ada kemungkinan tidak menemukan kombinasi hyperparameter yang optimal.

   c) **Cross-validation**
      Cross-validation digunakan untuk mengevaluasi kinerja model dengan membagi data menjadi beberapa bagian dan melatih model pada subset yang berbeda.
      - Trade-off:
        Lebih komputasi-intensif, tetapi memberikan evaluasi yang lebih stabil terhadap generalisasi model.

5. Langkah Mengatasi Residual Plot dengan Pola Non-linear dan Heteroskedastisitas
   a) Transformasi Data
      - Log Transform
        Jika residual plot menunjukkan pola non-linear atau heteroskedastisitas, salah satu solusi adalah mentransformasikan variabel target atau prediktor dengan logaritma atau transformasi lainnya.

      - Efek pada Model
        Transformasi ini dapat membantu menstabilkan varians dan mengurangi masalah heteroskedastisitas.

   b) Perubahan Model
      - Model Non-Linear
        Jika regresi linear tidak memadai, pertimbangkan untuk menggunakan Decision Tree atau Random Forest, yang dapat menangani hubungan non-linear tanpa perlu transformasi data eksplisit.

      - Efek pada Model
        Model non-linear lebih fleksibel dan dapat menangani data dengan pola yang lebih kompleks.