Evaluation Metrics and Regression Implementation

Theoretical Answers

1. **R-squared** represents the proportion of variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1.

2. **Assumptions of Linear Regression**:
    - Linearity
    - Independence
    - Homoscedasticity (equal variance of errors)
    - Normality of residuals
    - No multicollinearity

3. **R-squared vs Adjusted R-squared**:
    - R^2 increases with more variables.
    - Adjusted R^2 adjusts for the number of predictors and only increases if the new variable improves the model.

4. **Mean Squared Error (MSE)** measures the average of the squares of the errors. It penalizes larger errors more than MAE.

5. **Adjusted $R^2 = 0.85$** indicates that 85% of the variation in the dependent variable is explained by the model, adjusted for the number of predictors.

6. **Check normality of residuals** using:
    - Histogram/QQ Plot
    - Shapiro-Wilk test

    - Kolmogorov–Smirnov test

7. **Multicollinearity** occurs when independent variables are highly correlated. It inflates the variance of coefficients and can make the model unstable.

8. **Mean Absolute Error (MAE)** is the average of the absolute differences between predictions and actual values.

9. **Benefits of ML pipeline**:

    - Streamlined workflow

    - Reproducibility

    - Automation of preprocessing and modeling

10. **RMSE > MSE** for interpretation as it is in the same unit as the target variable, making it easier to understand.

11. **Pickling in Python** serializes a Python object into a byte stream. It's useful for saving ML models.

12. **High R-squared** means the model explains a large portion of variance, but it doesn't guarantee good predictions.

13. **Violation of assumptions** can lead to biased, inefficient, or misleading estimates.

14. **Address multicollinearity** by:

    - Removing correlated features

    - Using PCA

    - Ridge regression

15. **Feature selection** removes irrelevant variables, improves model performance, and reduces overfitting.

16. **Adjusted R-squared formula**:
    Adjusted R2=1−((1−R2)(n−1)n−p−1)\text{Adjusted } R^2 = 1 - \left(\frac{(1 - R^2)(n - 1)}{n - p - 1} \right)Adjusted R2=1−(n−p−1(1−R2)(n−1))
    where nnn = number of observations, ppp = number of predictors.

17. **MSE sensitivity to outliers**: Large errors are squared, so outliers heavily influence the value.

18. **Homoscedasticity** means constant variance of errors. It's essential for valid statistical inference.

19. **RMSE** is the square root of MSE and provides error in the same units as the target variable.

20. **Risk of pickling**: It can execute arbitrary code during unpickling, leading to security issues.

21. **Alternatives to pickling**:

    - `joblib`
    - `ONNX`
    - `HDF5` for models like Keras

22. **Heteroscedasticity**: Non-constant variance of residuals. It invalidates statistical tests and makes inference unreliable.

23. **Interaction terms** allow the model to capture combined effects of features, enhancing predictive power.