Evaluation Metrics and Regression:

ASSIGNMENT:

1. What does R-squared represent in a regression model?

- R-squared (also known as the coefficient of determination) represents the proportion of the variance in the dependent variable (the variable you're trying to predict) that is predictable from the independent variables (the variables you're using to make predictions).
- In simpler terms, it tells you how well the regression model fits the observed data.
- It ranges from 0 to 1, where:
    - 0 indicates that the model explains none of the variance.
    - 1 indicates that the model perfectly explains all of the variance.

2. What are the assumptions of linear regression?

- Linearity: The relationship between the independent and dependent variables is linear.
- Independence: The residuals (the differences between the observed and predicted values) are independent of each other.
- Homoscedasticity: The residuals have constant variance across all levels of the independent variables.
- Normality: The residuals are normally distributed.
- No multicollinearity: The independent variables are not highly correlated with each other.

3. What is the difference between R-squared and Adjusted R-squared?

- R-squared: Increases as you add more independent variables to the model, even if those variables don't actually improve the model's fit.
- Adjusted R-squared: Penalizes the model for adding unnecessary independent variables. It adjusts the R-squared value based on the number of independent

variables and the sample size. Therefore it is a better metric when comparing models with different numbers of independent variables.
- Adjusted R-squared will increase only when a new variable improves the model more than would be expected by chance.

4. Why do we use Mean Squared Error (MSE)?

- MSE measures the average squared difference between the predicted and actual values.
- It penalizes larger errors more heavily due to the squaring, making it sensitive to outliers.
- It's used because it's mathematically convenient and differentiable, which is important for optimization algorithms used in regression.

5. What does an Adjusted R-squared value of 0.85 indicate?

- An Adjusted R-squared of 0.85 means that approximately 85% of the variance in the dependent variable is explained by the independent variables in the model, after accounting for the number of predictors.
- This indicates a strong fit of the model to the data.

6. How do we check for normality of residuals in linear regression?

- Visual methods:
    - Histograms: Check if the residuals form a bell-shaped curve.
    - Q-Q plots (quantile-quantile plots): Check if the residuals fall along a straight line.
- Statistical tests:
    - Shapiro-Wilk test
    - Kolmogorov-Smirnov test

7. What is multicollinearity, and how does it impact regression?

- Multicollinearity occurs when two or more independent variables in a regression model are highly correlated.
- Impact:
    - Makes it difficult to determine the individual effect of each independent variable on the dependent variable.
    - Increases the variance of the regression coefficients, making them unstable.

    - Can lead to inflated standard errors, which can result in statistically insignificant coefficients.

8. What is Mean Absolute Error (MAE)?

- MAE measures the average absolute difference between the predicted and actual values.
- It gives equal weight to all errors, regardless of their magnitude.
- Provides a more robust measure of error in the presence of outliers compared to MSE.

9. What are the benefits of using an ML pipeline?

- Streamlines the workflow: Combines data preprocessing, feature engineering, and model training into a single, cohesive unit.
- Improves reproducibility: Ensures consistent application of preprocessing steps.
- Simplifies model deployment: Makes it easier to deploy the trained model.
- Reduces errors: Minimizes the risk of data leakage and other common mistakes.

10. Why is RMSE considered more interpretable than MSE?

- RMSE (Root Mean Squared Error) is the square root of MSE.
- Because MSE is squared, the units of the error are also squared. RMSE returns the error back to the original units of the dependent variable, making it easier to interpret. For example, if you are predicting home prices, RMSE will be in dollars, where as MSE will be Dollars squared.
- It provides a measure of the average magnitude of the errors in the same units as the target variable.

11. What is pickling in Python, and how is it useful in ML?

- Pickling is the process of serializing a Python object into a byte stream.
- In ML, it's used to save trained models, allowing you to load and use them later without retraining. This is useful for model deployment.

12. What does a high R-squared value mean?

- A high R-squared value (closer to 1) indicates that a large proportion of the variance in the dependent variable is explained by the independent variables.
- It suggests that the model is a good fit for the data.

- However, a high R-squared doesn't necessarily mean the model is perfect or that the independent variables are causally related to the dependent variable.

13. What happens if linear regression assumptions are violated?

- Biased or inefficient estimates: The regression coefficients may not be accurate.
- Invalid statistical inferences: The p-values and confidence intervals may be incorrect.
- Reduced model reliability: The model may not generalize well to new data.

14. How can we address multicollinearity in regression?

- Remove one of the correlated variables.
- Use dimensionality reduction techniques (e.g., principal component analysis).
- Combine the correlated variables into a single variable.
- Use regularization techniques (e.g., Ridge or Lasso regression).

15. How can feature selection improve model performance in regression analysis?

- Reduces overfitting: By removing irrelevant or redundant features, the model becomes simpler and less prone to overfitting.
- Improves model interpretability: Makes the model easier to understand and explain.
- Reduces computational cost: Simplifies the model and speeds up training and prediction.
- Improve model accuracy: by removing noise from the data.

16. How is Adjusted R-squared calculated?

- Adjusted $R^2 = 1 - [(1 - R^2) * (n - 1) / (n - k - 1)]$
    - Where:
        - R^2 is the R-squared value.
        - n is the number of observations.
        - k is the number of independent variables.

17. Why is MSE sensitive to outliers?

- MSE squares the errors, so large errors (outliers) have a disproportionately large impact on the overall MSE value.

18. What is the role of homoscedasticity in linear regression?

- Homoscedasticity ensures that the variance of the residuals is constant across all levels of the independent variables.
- This is important for accurate statistical inference, as it ensures that the standard errors of the regression coefficients are reliable.

19. What is Root Mean Squared Error (RMSE)?

- RMSE is the square root of the MSE.
- It provides a measure of the average magnitude of the errors in the same units as the dependent variable.

20. Why is pickling considered risky?

- Security risks: Pickled files can execute arbitrary code when loaded, posing a security vulnerability if the file is from an untrusted source.
- Version incompatibility: Pickled objects may not be compatible across different versions of Python or libraries.

21. What alternatives exist to pickling for saving ML models?

- Joblib: A library that provides more efficient pickling and unpickling of NumPy arrays and other large objects.
- ONNX (Open Neural Network Exchange): An open standard for representing machine learning models, allowing them to be exchanged between different frameworks.
- Saving models in formats specific to ML frameworks (e.g., TensorFlow SavedModel, PyTorch state_dict).

22. What is heteroscedasticity, and why is it a problem?

- Heteroscedasticity occurs when the variance of the residuals is not constant across all levels of the independent variables.
- Problems:
    - invalidates statistical test of significance.
    - The standard errors of the regression coefficients are biased.
    - Can lead to inaccurate confidence intervals and p-values.
    - Can lead to inefficiency of the model.

23. How can interaction terms enhance a regression model's predictive power?

- Interaction terms allow you to model situations where the effect of one independent variable on the dependent variable depends on the level of another independent variable.
- This can capture more complex relationships between the variables and improve the model's accuracy.