**EVALUATION METRICS THEOROTICAL ASSIGNMENT**

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

R-squared, also known as the **coefficient of determination**, is a statistical measure used to evaluate how well a regression model explains the variability in the dependent variable. It indicates the proportion of the variance in the dependent variable that is predictable from the independent variables.

1. **What are the assumptions of linear regression?**

Linear regression relies on several assumptions to ensure that the results are valid and reliable. Violating these assumptions can lead to biased estimates, incorrect conclusions, or inefficiencies in predictions.

**Assumptions of Linear Regression are,**

* **Linearity**: The relationship between independent variables and the dependent variable is linear.
* **Independence**: Observations are independent; no autocorrelation.
* **Homoscedasticity**: The variance of residuals is constant across all levels of the independent variables.
* **Normality**: Residuals are approximately normally distributed.
* **Zero Mean of Errors**: Residuals have an average of zero.
* The features should be least related.

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

R-squared (R2R^2R2) and Adjusted R-squared are both statistical metrics used to evaluate the performance of regression models, but they serve slightly different purposes and have distinct characteristics:

**1. R-squared:** Measures the proportion of variance in the dependent variable that is explained by the independent variables in the model.

Formula : r squared = 1 – (SSR/SST)

* Values range from 0 to 1 (or 0% to 100%).
* A higher r square indicates that more variance is explained by the model.
* r square increases or stays the same when more predictors are added, even if those predictors are irrelevant

### ****2. Adjusted R-squared**:** Adjusted R-squared accounts for the number of predictors in the model and adjusts r square to penalize the addition of non-informative predictors.

**Formula: Adjusted r square = 1 – (((1- r square)(n-1))/(n-p-1))**

* Unlike r square, it can decrease if new predictors don't improve the model's performance significantly.
* Adjusted r square provides a more realistic assessment of model quality when there are multiple predictors.
* It can be negative if the model performs worse than a model with no predictors.

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

We use **Mean Squared Error (MSE)** in regression analysis because it is a reliable and widely accepted metric for evaluating the accuracy of a model's predictions. MSE provides a quantitative measure of how well the model's predicted values match the actual observed values by calculating the average squared difference between them.

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

An **Adjusted R-squared** value of **0.85** indicates that **85% of the variance in the dependent variable** (target variable) is explained by the independent variables (predictors) in the regression model.

The model effectively explains a significant portion (85%) of the variation in the target variable, which implies a strong relationship between the predictors and the target.

A value of 0.85 suggests a **highly predictive model** but does not necessarily mean the model is perfect. There is still 15% of the variance in the dependent variable that is not explained by the model.

An Adjusted R-squared of 0.85 is a strong indicator of a well-performing regression model, assuming the data and assumptions of the model are sound.

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

#### **Histogram of Residuals**

* Plot a histogram of the residuals.
* A bell-shaped curve indicates normality.

#### **Density Plot**

* Overlay a kernel density estimate (KDE) of residuals with a normal distribution curve.
* Significant deviations indicate non-normality.

**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, meaning they convey redundant or overlapping information. As a result, it becomes challenging to isolate the individual effect of each predictor on the dependent variable.

**Impacts of Multicollinearity on Regression**

1. **Unstable Coefficients**:
   * Regression coefficients become sensitive to small changes in data.
   * Signs and magnitudes of coefficients may flip unexpectedly.
2. **Reduced Interpretability**:
   * It becomes difficult to determine the individual effect of predictors since they are not independent.
3. **Inflated Standard Errors**:
   * Standard errors of coefficients increase, making them less reliable.
   * This leads to wider confidence intervals and higher p-values, reducing the likelihood of detecting significant predictors.
4. **Reduced Model Stability**:
   * Predictions might remain accurate, but the model becomes unstable, particularly when extrapolating to new data.

**8)What is Mean Absolute Error (MAE)?**

**Mean Absolute Error (MAE) :** Mean Absolute Error (MAE) is a metric used to measure the average magnitude of errors between predicted and actual values in regression models. It calculates the absolute difference between the predicted y^ and yi actual values, without considering the direction of the error.

**Formula**:

**9)What are the benefits of using an ML pipeline?**

**Benefits of ML Pipelines:**

* **Automation**: Automates repetitive tasks like data preprocessing, model training, and evaluation, ensuring consistent execution and reducing manual effort.
* **Modularity**: Breaks the process into reusable components (e.g., data cleaning, feature extraction), making it easier to update or swap parts without affecting the entire workflow.
* **Scalability**: Pipelines efficiently handle large datasets and scale using distributed computing frameworks, allowing for faster training on big data.
* **Collaboration**: Standardizes workflows, making it easier for teams to work together on the same project by following a defined process, improving coordination.
* **Enhanced Experimentation**: Simplifies testing and comparing different models, hyperparameters, and configurations, speeding up the experimentation and model selection process.
* **Seamless Deployment**: Ensures that the steps used during training are replicated in production, reducing discrepancies between development and deployment environments.
* **Version Control**: Tracks changes to data, models, and code, ensuring reproducibility of results and providing an audit trail for compliance and governance.

**10)Why is RMSE considered more interpretable than MSE?**

* **RMSE** is more interpretable because it gives the error in the same units as the original data, making it easier to understand and communicate.
* **RMSE** represents the average size of the errors in the same scale as the data, making it easier to understand how far off, on average, the predictions are from the actual values.

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

**Pickling** is the process of converting a Python object (such as a model, dictionary, or list) into a byte stream, which can then be saved to a file. This process is done using the pickle module in Python.

**How Pickling is Useful in Machine Learning (ML):**

* **Model Persistence**:
  + **Pickling** allows you to save a trained machine learning model to a file, so you don't need to retrain the model every time you run your code. You can load the pre-trained model whenever needed for predictions.
  + This is useful for deploying models to production or sharing models across different environments.
* **Efficiency**:
  + Saving and loading models using pickling is faster compared to retraining from scratch, saving both time and computational resources, especially for large models.
* **Sharing Models**:
  + Pickled files can be shared across different systems, allowing others to use the model without needing the original code or training data.
* **Versioning**:
  + You can save different versions of models using pickling and load a specific version based on requirements, facilitating model version control.

**12)What does a high R-squared value mean?**

A **high R-squared value** (close to 1) means that the model **explains a large proportion** of the variance in the dependent variable. In other words, the model's predictions are closely aligned with the actual data.

**13)What happens if linear regression assumptions are violated?**

If the assumptions of linear regression are violated, the results of the model can become unreliable, leading to incorrect interpretations, biased estimates, and misleading predictions

**Consequences of Violating Assumptions:**

1. **Bias**: Estimates of the regression coefficients can become biased, leading to incorrect predictions.
2. **Inefficiency**: The model may become inefficient, meaning it doesn't make the best use of the data.
3. **Invalid Inference**: Statistical tests, such as significance tests for the coefficients, can produce misleading results (wrong p-values or confidence intervals).
4. **Poor Generalization**: The model might overfit the training data and perform poorly on unseen data.

**14)How can we address multicollinearity in regression?**

We can address multicollinearity by:

* Remove or combine correlated variables.
* Use techniques like **PCA** or **regularization** (Ridge, Lasso).
* Consider increasing the sample size or using domain knowledge to select key features. By applying these strategies, you can improve the stability, interpretability, and predictive power of your regression model.

**15)Why do we use pipelines in machine learning?**

We use **pipelines** in machine learning to streamline the entire process, from data preprocessing to model deployment, ensuring efficiency, consistency, and reproducibility.

* Pipelines automate repetitive tasks such as data cleaning, feature selection, model training, and evaluation.
* Pipelines ensure that the same steps are executed in the same order every time, which helps maintain consistency across experiments and deployments
* A pipeline breaks the machine learning process into discrete, manageable steps (e.g., data preprocessing, model training, evaluation), making complex workflows easier to manage and debug.
* Pipelines are often designed to handle larger datasets efficiently and can be scaled to process more data or to run across distributed systems.
* Pipelines help track changes in the data, model, and parameters, making it easier to audit and version different stages of the machine learning process.
* Pipelines ensure that the same preprocessing steps used during training are applied during inference, ensuring consistency between training and deployment phases.

**16)How is Adjusted R-squared calculated?**

**Adjusted R-squared** is a modified version of R-squared that adjusts for the number of predictors in the model. It accounts for the fact that adding more variables to a regression model can artificially inflate R-squared, even if those variables don't contribute meaningfully to the model.

**Adjusted r square = 1 – (((1- r square)(n-1))/(n-p-1))**

**Adjustment**: The formula penalizes the addition of irrelevant predictors. As more variables are added, the denominator increases, which reduces the value of Adjusted R-squared if the added variables do not improve the model significantly.

**17) Why is MSE sensitive to outliers?**

Mean Squared Error (MSE) is particularly sensitive to outliers because it squares the difference between predicted values and actual values. This squaring function amplifies larger errors more significantly than smaller ones.

**18) What is the role of homoscedasticity in linear regression?**

Homoscedasticity is a key assumption in linear regression that refers to the consistency of variance in the errors across all levels of the independent variable(s).

Few roles of homoscedasticity are,

* **Reliable Coefficient Estimates**: Ensures unbiased and efficient coefficient estimates.
* **Accurate Standard Errors**: Provides reliable standard errors for confidence intervals and hypothesis testing.
* **Model Validity**: Indicates that the model is properly specified and fits the data well.
* **Predictive Performance**: Enhances the reliability of predictions for new data points.
* **Graphical Analysis**: Results in residual plots that show no patterns, confirming model assumptions are met.

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

**Root Mean Squared Error (RMSE)** is a metric used to measure the average magnitude of the errors between predicted values and actual values in a regression model. It is calculated by taking the square root of the average of the squared differences between these values. RMSE provides an indication of the accuracy of the model's predictions.

**20) Why is pickling considered risky?**

**pickling** can be considered risky for the following reasons:

* **Security Vulnerabilities**: Pickle files can execute arbitrary code during unpickling. If a pickle file is tampered with by a malicious actor, it could run harmful code on your system.
* **Compatibility Issues**: Pickle files are specific to the Python version and library versions used to create them. Unpickling a file created with different versions can lead to errors or unexpected behaviour.
* **Data Corruption**: Pickle files can become corrupted, leading to data loss. Unlike text formats like JSON or CSV, corrupted pickle files are harder to diagnose and recover.
* **Lack of Portability**: Pickle files are not human-readable and are not suitable for cross-language or cross-platform data sharing.

**21) What alternatives exist to pickling for saving ML models?**

* **TensorFlow SavedModel**: If you are using TensorFlow, the SavedModel format is a comprehensive, portable, and recoverable format for TensorFlow models.

Usage: model.save('path\_to\_my\_model')

* **Keras H5 Format**: Specifically for Keras models, the H5 format is handy and straightforward to use.

Usage: model.save('my\_model.h5')

* **JSON**: For simple model configurations and parameters, JSON can be used to serialize the model architecture and weights separately.

Usage: model.to\_json()and then saving weights with model.save\_weights('path\_to\_weights.h5')

* **Joblib**: This library, also in the Python ecosystem, is optimized for saving large numerical arrays and is more efficient than pickle for big data.

Usage: from joblib import dump, load

**22) What is heteroscedasticity, and why is it a problem?**

Heteroscedasticity refers to the situation in regression analysis where the variance of the errors (residuals) is not constant across all levels of the independent variable(s). In simpler terms, it means that the spread of the residuals changes at different levels of the independent variable, rather than remaining constant.

problems associated with heteroscedasticity are:

* **Biased Standard Errors**: Leads to incorrect estimates of standard errors, affecting confidence intervals and hypothesis tests.
* **Inefficient Estimates**: Ordinary Least Squares (OLS) estimates lose their efficiency and may no longer be the Best Linear Unbiased Estimators (BLUE).
* **Misleading Inferences**: Inaccurate p-values and t-statistics can lead to incorrect conclusions about the significance of predictors.
* **Model Validity**: Indicates that the model may be improperly specified or that important variables are missing.
* **Predictive Performance**: Reduces the reliability of predictions, as the variability in errors suggests the model isn't capturing all relevant information.

**23) How does adding irrelevant predictors affect R-squared and Adjusted R-squared?**

Adding irrelevant predictors affects R-squared and Adjusted R-squared in different ways:

### ****R-squared****:

* **Increases or Remains the Same**: Adding predictors, even if they are irrelevant, will either increase or not change the R-squared value. This is because R-squared measures the proportion of variance explained by the model, and with more predictors, the model tends to fit the data better, even if those predictors do not actually improve the model.

### ****Adjusted R-squared****:

* **Decreases or Remains the Same**: Adjusted R-squared, on the other hand, accounts for the number of predictors in the model. If you add irrelevant predictors, it will decrease or remain the same. This is because Adjusted R-squared penalizes the addition of variables that do not improve the model's explanatory power, providing a more accurate measure of model fit.