**PW\_REGRESSION\_ASSIGNMENT 2:**

**Q1. Explain the concept of R-squared in linear regression models. How is it calculated, and what does it represent?**

**Answer:**

R-squared (R² or the coefficient of determination) is a statistical measure in a regression model that determines the proportion of variance in the dependent variable that the independent variable can explain. In other words, R-squared shows how well the data fit the regression model (the goodness of fit).

R-squared = SS regression / SS total

Where, SS regression is the sum of squares due to regression (explained sum of squares);

and SS total is the total sum of squares

The most common interpretation of r-squared is how well the regression model explains observed data. For example, an R-squared of 60% reveals that the regression model explains 60% of the variability observed in the target variable. Generally, a higher R-squared indicates more variability is explained by the model.

**Q2. Define adjusted R-squared and explain how it differs from the regular R-squared.**

**Answer:**

Adjusted R-squared is a modified version of R-squared that corrects for overestimation of model accuracy. It's used to determine the goodness of fit for linear models and multiple regression models. The main difference between adjusted R-squared and R-squared is that adjusted R-squared accounts for the number of predictors in the model, while R-squared does not:

* **R-squared**

Measures the proportion of variance in the dependent variable explained by the model's independent variable(s). R-squared increases or remains the same when new predictors are added to the model.

* **Adjusted R-squared**

Corrects for overestimation by penalizing the inclusion of irrelevant variables. Adjusted R-squared only increases when a new predictor improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected.

Adjusted R-squared is always less than or equal to R-squared. A value of 1 indicates a model that perfectly predicts values in the target field, while a value that is less than or equal to 0 indicates a model that has no predictive value.

**Q3. When is it more appropriate to use adjusted R-squared?**

**Answer:**

Adjusted R-squared is a modified version of R-squared, and it adjusts the number of predictors in the model.

Adjusted R-squared is more appropriate to use than R-squared while working with multiple regression models:

* **Model comparison**

Adjusted R-squared helps to choose the best-performing model when comparing models with different numbers of predictors.

* **Preventing overfitting**

Adjusted R-squared penalizes the inclusion of irrelevant predictors, so it ensures that only predictors that improve the model's performance are included.

* **More accurate view of correlation**

Adjusted R-squared considers how many independent variables are added to a model and which one helps to get a more accurate view of the correlation between variables.

* **Better measure of model fit**

Adjusted R-squared prevents you from being misled by a high R-squared value that's due to overfitting.

R-squared is only intended to work in a simple linear regression model with one explanatory variable.

**Q4. What are RMSE, MSE, and MAE in the context of regression analysis? How are these metrics calculated, and what do they represent?**

**Answer:**

**Mean Absolute Error (MAE)**

**MAE** treats absolute errors linearly - a change in the error will have a proportional effect on **MAE**. For example, an error of 40 is twice as bad as an error of 20.

However, to build models that don’t generate larger errors too often. Thus, **we need a metric that penalizes larger errors more harshly than smaller ones**.

A metric using the **square of errors can be created**. That’ll ensure that a larger error will produce a far more pronounced effect.

Consider two error values - 20 and 40. Their squared values are 400 and 1600, respectively. Even though 40 is twice of 20, it’ll contribute 4 times to the total squared error.

*Mean Absolute Error* (*MAE*) = (*Sum of Absolute Errors*​) / (*Number of Predictions)*

**Mean Squared Error (MSE)**

**MSE**is a helpful metric, but**it is hard to interpret.** It, by definition, involved squaring of error terms. Thus, **MSE** doesn’t have the same units as the value we want to predict.

**Root Mean Squared Error (RMSE)**

Taking a square root of MSE will give us Root Mean Squared Error (RMSE):

Root Mean Squared Error (RMSE) = Square root of MSE

**Q5. Discuss the advantages and disadvantages of using RMSE, MSE, and MAE as evaluation metrics in regression analysis.**

**Answer:**

|  |  |  |
| --- | --- | --- |
| Metric | Advantages | Disadvantages |
| MAE | Easy to interpret and understand. Less sensitive to outliers. | Does not consider the direction of errors. |
| MSE | Penalizes larger errors more heavily, giving it more sensitivity to outliers. | Harder to interpret that MAE, as it is not in the same unit as the original data. |
| RMSE | It has the same units as the original data, making it easier to interpret. | Sensitive to outliers. |

**Q6. Explain the concept of Lasso regularization. How does it differ from Ridge regularization, and when is it more appropriate to use?**

**Answer:**

Lasso regression, commonly referred to as L1 regularization, is a method for stopping overfitting in linear regression models by including a penalty term in the cost function. In contrast to Ridge regression, it adds the total of the absolute values of the coefficients rather than the sum of the squared coefficients.

Lasso regression can reduce certain coefficients to zero, conducting feature selection in effect. With high-dimensional datasets where many characteristics could be unnecessary or redundant, this is very helpful. The resultant model is less complex and easier to understand, and by minimizing overfitting, it frequently exhibits improved predictive performance.

**Q7. How do regularized linear models help to prevent overfitting in machine learning? Provide an example to illustrate.**

**Answer:**

Regularization reduces overfitting by penalizing larger weights, encouraging the model to prioritize simpler hypotheses. Improving Model Generalization: Regularization helps ensure the model performs well on the training and new, unseen data by constraining its complexity.

Regularization improves machine learning models by correcting overfitting and enabling them to generalize on new, unseen data.

**Q8. Discuss the limitations of regularized linear models and explain why they may not always be the best choice for regression analysis.**

**Answer:**

The limitations of Regularization are:

* Regularization can make the model too simple and unsuitable for the data if the regularization parameter is too large. ...
* Regularization assumes that all input variables are of equal importance, which may not be true in some cases.

Given that it presumes a linear relationship between the input and output variables, linear regression is unable to accurately fit in complicated datasets. Since the relationships between the dataset's variables are rarely linear in real-world situations, a straight line cannot accurately represent the data.

**Q9. You are comparing the performance of two regression models using different evaluation metrics.**

**Model A has an RMSE of 10, while Model B has an MAE of 8. Which model would you choose as the better performer, and why? Are there any limitations to your choice of metric?**

**Answer:**

Both RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) are popular metrics for evaluating the performance of regression models, but they emphasize different aspects of prediction accuracy. Let’s analyze the situation using both metrics:

Model A: RMSE = 10  
Model B: MAE = 8

The choice of which model is better depending on the specific context and the importance of different aspects of prediction accuracy:

- RMSE: RMSE gives more weight to larger errors due to the squaring operation. This can be advantageous when significant errors are particularly undesirable. However, RMSE is sensitive to outliers and can be influenced by extreme values, which might not reflect the overall model performance accurately.

* MAE: MAE represents the average magnitude of errors, giving equal weight to all errors. It is less sensitive to outliers and can provide a more robust assessment of overall prediction accuracy.

Comparing the two models:

* Model A has a lower RMSE (10), indicating that it might perform better when large errors are of particular concern.
* Model B has a lower MAE (8), suggesting that it might perform better in terms of overall average prediction accuracy.

Since both RMSE and MAE have their strengths and weaknesses, the choice of the better model depends on the specific goals of the analysis and the relative importance of minimizing large errors versus achieving a more balanced average prediction accuracy.

Limitations of Metric Choice:

1. Domain Context: The choice of metric should consider the context of the problem. Depending on the domain and the consequences of different types of errors, one metric might be more appropriate than the other.

2. Outlier Sensitivity: RMSE can be strongly affected by outliers due to the squaring operation. If the dataset contains outliers, RMSE might overemphasize their impact.

3. Scale Considerations: RMSE and MAE are in different units than the original variable. Comparing models based solely on these metrics might not provide the full picture if the scales of the variables are different.

4. Model Goal: The goal of the model also matters. If the model is intended for decision-making, cost considerations, or specific accuracy requirements, the choice of metric should align with those goals.

5. Model Complexity: Metrics should be interpreted alongside model complexity. A simpler model with slightly higher errors might still be preferable due to its interpretability and generalization.

**Q10. You are comparing the performance of two regularized linear models using different types of regularization. Model A uses Ridge regularization with a regularization parameter of 0.1, while Model B uses Lasso regularization with a regularization parameter of 0.5. Which model would you choose as the better performer, and why? Are there any trade-offs or limitations to your choice of regularization method?**

**Answer:**

Ridge Regression with a regularization parameter of 0.5 tends to perform better when many predictors contribute to the outcome, as it shrinks coefficients but retains all variables. This can lead to lower variance in predictions, especially in high-dimensional spaces, but may not simplify the model significantly.

Limitations of Ridge and Lasso Regressions:

The Ridge-Lasso approach is limited in requiring the input features to be standardized before fitting the model. It means that any feature with a large range of values can bias results because of its scale relative to other features with smaller ranges.

\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*