Advanced Regression Assignment Part - II

Q-1:

What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

A-1:

The optimal value of alpha for ridge and lasso regression depends on the dataset and goals of the modeling process. It is determined by hyperparameter tuning. Cross-Validation techniques often employed to select the optimal alpha value which minimizes the prediction error. K-Fold cross validation is one such technique.
Higher the value of alpha more the regularization.
If the value of alpha is doubled, following changes will occur

- Ridge Regression:
  - With the larger values of alpha, the amount of L2 regularization effect in the Ridge model increases and the model's complexity decreases.
- Lasso Regression:
  - Larger values of alpha in Lasso Regression will increase the amount of L1 regularization effect.
  - As the alpha increases, more predictor variables will become zero and excluded from the model.

After the value of alpha is doubled, the predictors which do not become zero are the most important ones and to be used for prediction.

Q-2:

You have determined the optimal value of lambda for ridge and lasso regression during the assignment. Now, which one will you choose to apply and why?

A-2:

- For ridge model r2 score was 93.63% for training data and 88.31% for test data.
- For lasso model r2 score was 91.00% for training data and 88.37% for test data.
- Both the models have very good training and test scores. But the Lasso model has a narrower difference (2.63%) between the train score and test score compared to the Ridge Model (5.32%).
- Also Lasso shrieked many predictors to zero hence the number of significant parameters are lesser compared to Ridge Model.

- So I would choose to apply Lasso compared to Ridge.

Q-3:

After building the model, you realized that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

A-3:

After building another model excluding the five predictors, below listed steps to be followed to get five most important predictor variables.

- **Step 1**: Retrieve coefficient values of predictors from the model. Larger the coefficient, the more important the predictor.
- **Step 2**: Sort these values in descending order so that most important predictors are listed first.
- **Step 3**: Select top 5 predictors from the sorted list.

The selected 5 predictors are now the most important predictor variables.

Q-4:

How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

A-4:

To ensure that a model is robust and generalizable one can practice below listed steps:

1. **Feature Selection and Engineering:**
    a. Select relevant features that contribute to the predictive power of the model.
    b. Engineer new features that might improve model performance.
    c. Avoid overfitting by selecting only necessary features or using regularization techniques.
2. **Cross-Validation:**
    a. Use cross-validation techniques such as k-fold cross-validation to assess model performance on multiple subsets of the data.
    b. Evaluate the model's performance metrics (e.g., accuracy, precision, recall, F1-score) across different folds to ensure consistency.
3. **Hyperparameter Tuning:**

    a. Tune model hyperparameters using techniques like grid search, randomized search, or Bayesian optimization.
4. **Regularization:**
    a. Apply regularization techniques such as L1 or L2 regularization to prevent overfitting and encourage model simplicity.
5. **Model Interpretability:**
    a. Ensure that the model is interpretable, especially in domains where interpretability is crucial.
    b. Use techniques such as feature importance analysis or model-agnostic methods for explaining predictions.
6. **Regular Monitoring:**
    a. Continuously monitor model performance in production. Models can degrade over time due to changing data distributions, so it's essential to re-evaluate and retrain them periodically.

Implications of robustness and generalisability for the accuracy of the model and why?

Ensuring robustness and generalizability in a machine learning model requires a trade-off with model accuracy. Here are some implications of focusing on robustness and generalizability on model accuracy:

- **Slightly Reduced Accuracy:** When focus is on prioritizing robustness and generalizability, model's accuracy will be slightly compromised. Techniques such as regularization, feature selection, cross-validation etc. prevent the model from overfitting. Such a model may not fit well the training data leading to a decrease in accuracy.
- **Better prediction on Unseen Data:** Focusing on robustness and generalizability often leads to a model which learns to capture the underlying patterns in the training data. This results in a model better at generalizing new and unseen data. Such models are better than highly accurate models.
- **Better Handling of Outliers:** Models trained with an emphasis on robustness are better equipped to handle outliers or anomalies in the data. Instead of being overly influenced by extreme data points, these models learn to generalize patterns more effectively, resulting in more consistent predictions across different data distributions.
- **Interpretability and Explainability:** Prioritizing robustness and generalizability often leads to simpler, more interpretable models. Complex models with high accuracy are harder to interpret and explain compared to simpler models.