Assignment-based Subjective Questions

Question 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?

Answer –

1. The optimal value of alpha for Ridge regression is 10.0 and Lasso regression is 100.0.
2. If we double the value of alpha for both Ridge and Lasso regression, the model will become more biased and have lower variance. This will lead to: Shrinkage of the coefficient estimates towards zero. Less complexity in the models, as more coefficients will be reduced to zero. Lower risk of overfitting, but with the potential for underfitting if the value of alpha is too high.
3. To determine the most important predictor variables after the change is implemented, you can look at the absolute values of the coefficients in the models. The predictors with higher absolute coefficient values will be more influential.
    - In the case of Ridge regression, all coefficients will be non-zero, but some will be closer to zero than others.
    - In Lasso regression, some coefficients will be exactly zero because Lasso regression has a built-in feature selection mechanism.

Question 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?

Answer –

The value of alpha in Lasso regression controls the amount of regularization applied to the model. A higher value of alpha implies stronger regularization, which can lead to better generalization performance and reduced risk of overfitting.

Furthermore, Lasso regression has the added benefit of performing feature selection, as it sets the coefficients of less important features to zero.

Difference between Lasso Regression and Ridge Regression is as follows –

S.NO.	L1 Regularization (Lasso Regression)	L2 Regularization (Ridge Regression)
1.	Lasso Regression adds “absolute value of magnitude” of coefficient as penalty term to the loss function	Ridge regression adds “squared magnitude” of coefficient as penalty term to the loss function.
2.	Lasso regression can produce many solutions to the same problem.	Ridge regression can only produce one solution to one problem
3.	Lasso regression usually produces unstable solutions which means the regression line may be uneven and jump by a large amount	Ridge Regression produces stable solutions which means that even for a small adjustment to a data point, the regression line will not move too much
4.	Lasso regression has built in feature selection	Ridge regression does not have built in feature selection
5.	Lasso regression produces sparse outputs	Ridge regression produces non sparse outputs
6.	For Lasso regression, the computation is inefficient for non-sparse cases.	For ridge regression, the computation is efficient because due to analytical solutions

Lasso regression would be a better option it would help in feature elimination and the model will be more robust.

Question 3 – After building the model, you realised 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?

<u>Answer –</u>

The five most important predictor variables. Here are the steps:

- We can do this by checking the absolute values of the coefficients or feature importances, depending on the library you used to build the model.
- After identifying the five most important predictor variables, exclude them from the original dataset.
- Preprocess the data (one-hot encoding, scaling, etc.) as necessary for the remaining predictor variables.
- Split the pre-processed data into training and testing sets.
- Train a new Lasso model using the pre-processed data with the five most important predictor variables excluded.

Question 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?

Answer –

To ensure that a model is robust and generalizable, following are the best practices:

- Data Diversity: Collect data from various sources, locations, or time frames to capture the variability in the data and make the model more robust.
- Cross-validation: Divide the data into multiple folds for training and testing. This technique helps in evaluating the model performance across different subsets, reducing the chance of overfitting and making the model more robust.
- Regularization: Apply techniques such as L1 (Lasso) or L2 (Ridge) regularization to penalize large coefficients and prevent overfitting, making the model more generalizable.
- Hyperparameter Tuning: Optimize the model's hyperparameters using grid search, random search, or Bayesian optimization to find the best combination of parameters for your problem.
- Feature Selection: Select relevant features to include in the model, which can help improve robustness and generalizability.
- Model Evaluation: Use appropriate evaluation metrics, such as R-squared, mean squared error, or cross-validation scores, to ensure that the model is robust and generalizable.

Implications of a robust and generalizable model:

- Better Accuracy: A robust and generalizable model tends to have better accuracy on unseen data compared to a model that is overfitted or underfitted.
- Lower Bias and Variance: Robust and generalizable models have a balance between bias and variance, reducing the chance of underfitting (high bias) and overfitting (high variance).
- Improved Decision Making: Models that are robust and generalizable provide more reliable predictions, resulting in better decision-making and predictions in real-world applications.

In summary, building a robust and generalizable model is crucial for obtaining accurate and reliable predictions. By following best practices such as data diversity, cross-validation, regularization, hyperparameter tuning, feature selection, and model evaluation, we can improve the robustness and generalizability of our model.