House Price Predictions – 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

As per the model, the optimal value of alpha for ridge and lasso regression is 4.0 and 100.0 respectively. We created the model with doubled the value of alpha for both ridge and lasso. Refer below table to understand the performance metrics We see that the performance has dropped a little with doubled alpha values for lasso and ridge values.

	Metric	Linear_Regression	L2-Ridge	L2-Ridge_Revised	L1-Lasso	L1-Lasso_Revised
0	R2 Score (Train)	8.422057e-01	8.392398e-01	8.344913e-01	8.383981e-01	8.322911e-01
1	R2 Score (Test)	8.283781e-01	8.277685e-01	8.244060e-01	8.283886e-01	8.221070e-01
2	RSS (Train)	1.006838e+12	1.025762e+12	1.056061e+12	1.183352e+16	1.170594e+16
3	RSS (Test)	4.837528e+11	4.854710e+11	4.949491e+11	2.108775e+15	2.079486e+15
4	MSE (Train)	3.140269e+04	3.169644e+04	3.216115e+04	3.177931e+04	3.237422e+04
5	MSE (Test)	3.323340e+04	3.329236e+04	3.361578e+04	3.323237e+04	3.383513e+04

The top important variables with new values of alpha are:

1. Neighborhood_NoRidge
2. Neighborhood_NridgHt
3. BsmtExposure_Gd
4. GrLivArea
5. OverallQual
6. Exterior1st_BrkFace
7. Neighborhood_Crawfor
8. Neighborhood_Somerst
9. Neighborhood_StoneBr
10. LandContour_Low

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 2

Based on the performance metrics below we will go ahead and choose Lasso Regression.

1. R-Square of Lasso Regularization is better.
2. Residual Sum of Squares (RSS) Error of Lasso Regularization is less.
3. Mean Square Error (MSE) Error of Lasso Regularization is less.

Refer below table for details

	Metric	Linear_Regression	L2-Ridge	L2-Ridge_Revised	L1-Lasso	L1-Lasso_Revised
0	R2 Score (Train)	8.422057e-01	8.392398e-01	8.344913e-01	8.383981e-01	8.322911e-01
1	R2 Score (Test)	8.283781e-01	8.277685e-01	8.244060e-01	8.283886e-01	8.221070e-01
2	RSS (Train)	1.006838e+12	1.025762e+12	1.056061e+12	1.183352e+16	1.170594e+16
3	RSS (Test)	4.837528e+11	4.854710e+11	4.949491e+11	2.108775e+15	2.079486e+15
4	MSE (Train)	3.140269e+04	3.169644e+04	3.216115e+04	3.177931e+04	3.237422e+04
5	MSE (Test)	3.323340e+04	3.329236e+04	3.361578e+04	3.323237e+04	3.383513e+04

Question 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?

Answer 3

To answer this, we dropped the five most important predictor variables and re-trained the model using Lasso Regularization. Based on new values of the coefficients, below are the next set of five most important predictor variables.

1. Exterior2nd_ImStucc
2. Neighborhood_Crawfor
3. Exterior1st_BrkFace
4. Neighborhood_StoneBr
5. LandContour_Low

Refer file “Housing_Price_Prediction.ipynb” for details.

Question 4

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

Answer 4

To ensure that a linear regression model with L1/L2 regularization is robust and generalizable, you can consider the following approaches:

1. **Model Complexity Evaluation:** Assess the complexity of your model by evaluating metrics such as the number of features, the number of non-zero coefficients (in the case of LASSO), or the model's effective degrees of freedom. Aim to strike a balance between model complexity and generalization performance.
2. **Handle model complexity**
    a. **Overfitting:** If the regularization parameter is too low, the model may overfit the training data, leading to poor generalization performance on unseen data.
    b. **Underfitting:** If the regularization parameter is too high, the model may become too simple and fail to capture the underlying patterns in the data, resulting in poor performance on both the training and unseen data. Feature Selection: In addition to regularization, consider performing feature selection to identify the most relevant features for your model. This can help reduce the model complexity and improve

generalization. Techniques like recursive feature elimination (RFE), LASSO regression, (or PCA) can be used for feature selection.

3. **Cross-Validation:** We can perform cross-validation to assess the model's performance on unseen data. This involves splitting the dataset into training and validation/test sets, training the model on the training set, and evaluating its performance on the validation/test set. Cross-validation helps you estimate the model's true generalization performance and identify any potential overfitting or underfitting issues.
4. **Regularization Hyperparameter Tuning:** Carefully tune the regularization hyperparameters (lambda for L2 regularization or alpha for L1 regularization) using techniques like grid search or cross-validation. The appropriate choice of regularization strength can help strike a balance between model complexity and generalization.
5. **Evaluation Metrics:** Use appropriate evaluation metrics to assess the model's performance, such as mean squared error (MSE), R-squared (R^2), or mean absolute error (MAE). These metrics can provide insights into the model's fit and predictive capabilities on both the training and validation/test sets.

Regarding the implications of these approaches for the model's accuracy:

Implementing the above techniques can help ensure that your linear regression model is robust and generalizable, meaning that it can perform well on unseen data, not just the training data. This is crucial for real-world applications where the model needs to make accurate predictions on new, unseen instances.

1. **Improved Accuracy:** A well-tuned and regularized model that is robust and generalizable is more likely to achieve higher accuracy on both the training and unseen data. Overfitting can lead to high training accuracy but poor performance on new data, while underfitting can result in low accuracy on both training and unseen data. The right balance of model complexity and regularization can help maximize the model's predictive performance.
2. **Reduced Overfitting:** Regularization techniques like L1 (LASSO) and L2 (Ridge) can help prevent overfitting by introducing a penalty for model complexity. This encourages the model to learn a simpler, more generalizable representation of the data, leading to better performance on unseen instances.
3. **Interpretability:** Depending on the type of regularization used, the model may become more interpretable. For example, LASSO (L1 regularization) can result in sparse models, where only the most important features have non-zero coefficients, making it easier to understand the model's decision-making process.

In summary, ensuring that your linear regression model is robust and generalizable through techniques like cross-validation, regularization hyperparameter tuning, and feature selection can significantly improve the model's accuracy and generalization performance, which is crucial for real-world applications.