Assignment 2

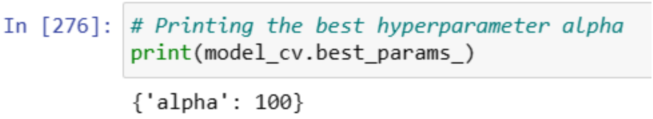
Question 1:

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:

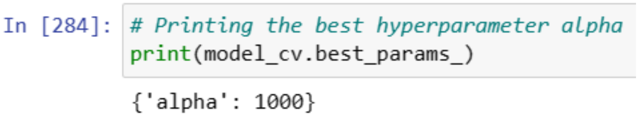
- Alpha is the regularization parameter that controls the amount of regularization applied in the models. Higher values of alpha result in stronger regularization, which can help prevent over fitting but may also lead to under fitting.

The optimal values of alpha for ridge are as below:



```
In [276]: # Printing the best hyperparameter alpha
          print(model_cv.best_params_)

{'alpha': 100}
```

The optimal values of alpha for Lasso are as below:



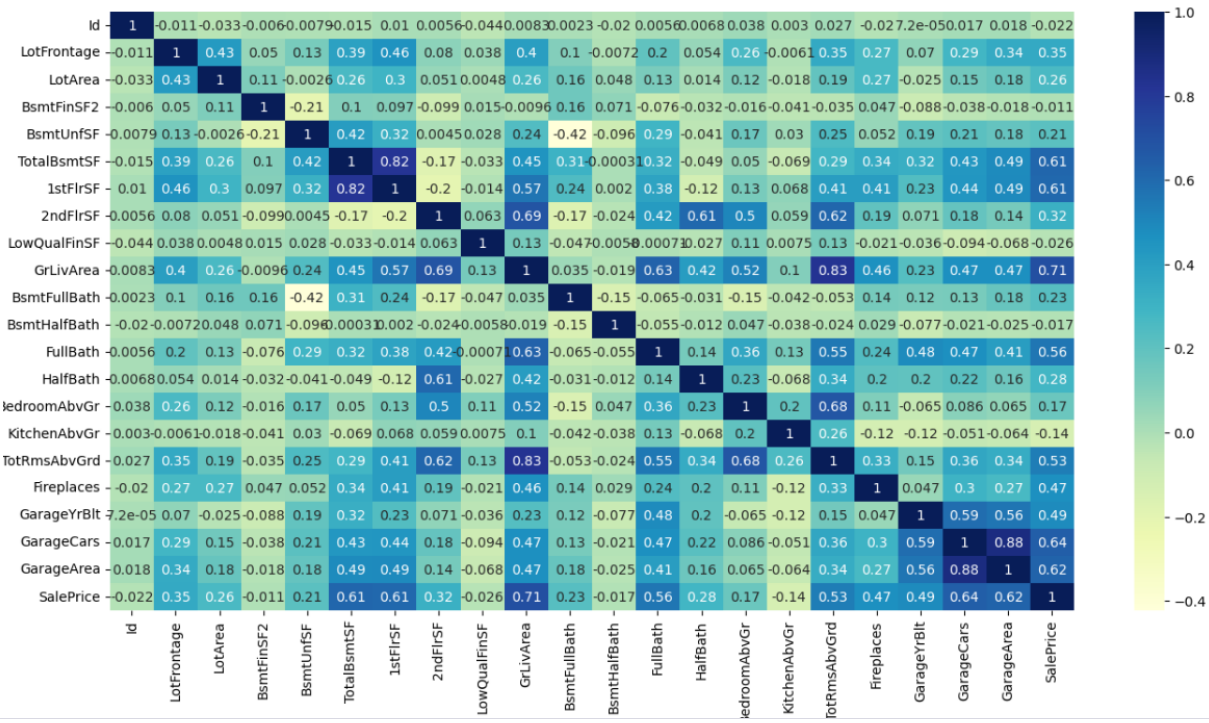
```
In [284]: # Printing the best hyperparameter alpha
          print(model_cv.best_params_)

{'alpha': 1000}
```

- If we double the value of alpha for both ridge and lasso regression, the models will be more heavily regularized. This means that the models will put stronger constraints on the coefficients and may shrink them further towards zero. Consequently, the impact on the model will be:

Ridge Regression: Increasing alpha in ridge regression will increase the amount of shrinkage applied to the coefficients. The coefficients will be further reduced towards zero but will not be completely eliminated. This can result in a simpler model with potentially fewer important predictors, as some coefficients may be further diminished or become close to zero.

Lasso Regression: Doubling alpha in lasso regression will lead to increased sparsity in the model. The higher regularization will push more coefficients to exactly zero, effectively removing them from the model. As a result, the number of important predictor variables will decrease, and the model will become more simplified with a reduced number of active predictors.

- The most important predictor variable is in data set ‘Saleprice’.



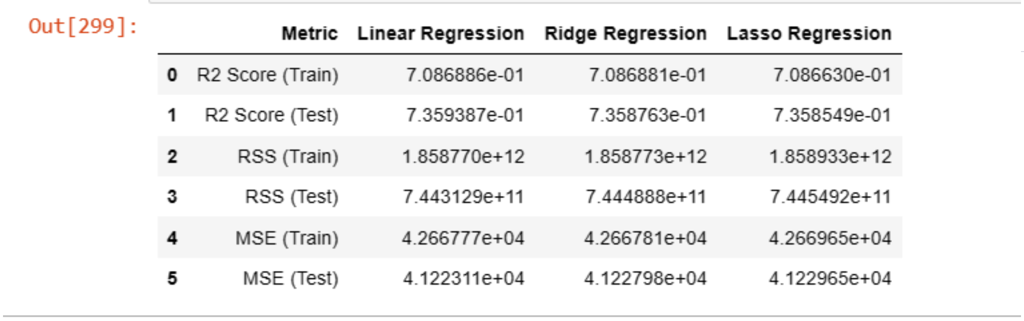
	Id	LotFrontage	LotArea	BsmtFinSF2	BsmtUnfSF	TotalBsmtSF	1stFlrSF	2ndFlrSF	LowQualFinSF	GrLivArea	BsmtFullBath	BsmtHalfBath	FullBath	HalfBath	edroomAbvGr	KitchenAbvGr	TotRmsAbvGrd	Fireplaces	GarageYrBlt	GarageCars	GarageArea	SalePrice
Id	1	-0.011	-0.033	-0.006	-0.0079	-0.015	0.01	0.0056	-0.044	0.0083	0.0023	-0.02	0.0056	0.0068	0.038	0.003	0.027	-0.02	7.2e-05	0.017	0.018	-0.022
LotFrontage	-0.011	1	0.43	0.05	0.13	0.39	0.46	0.08	0.038	0.4	0.1	-0.0072	0.2	0.054	0.26	-0.0061	0.35	0.27	0.07	0.29	0.34	0.35
LotArea	-0.033	0.43	1	0.11	-0.0026	0.26	0.3	0.051	0.0048	0.26	0.16	0.048	0.13	0.014	0.12	-0.018	0.19	0.27	-0.025	0.15	0.18	0.26
BsmtFinSF2	-0.006	0.05	0.11	1	-0.21	0.1	0.097	-0.099	0.015	-0.0096	0.16	0.071	-0.076	-0.032	-0.016	-0.041	-0.035	0.047	-0.088	-0.038	-0.018	-0.011
BsmtUnfSF	-0.0079	0.13	-0.0026	-0.21	1	0.42	0.32	0.0045	0.028	0.24	-0.42	-0.096	0.29	-0.041	0.17	0.03	0.25	0.052	0.19	0.21	0.18	0.21
TotalBsmtSF	-0.015	0.39	0.26	0.1	0.42	1	0.82	-0.17	-0.033	0.45	0.31	-0.00031	0.32	-0.049	0.05	-0.069	0.29	0.34	0.32	0.43	0.49	0.61
1stFlrSF	0.01	0.46	0.3	0.097	0.32	0.82	1	-0.2	-0.014	0.57	0.24	0.002	0.38	-0.12	0.13	0.068	0.41	0.41	0.23	0.44	0.49	0.61
2ndFlrSF	0.0056	0.08	0.051	-0.099	0.0045	-0.17	-0.2	1	0.063	0.69	-0.17	-0.024	0.42	0.61	0.5	0.059	0.62	0.19	0.071	0.18	0.14	0.32
LowQualFinSF	-0.044	0.038	0.0048	0.015	0.028	-0.033	-0.014	0.063	1	0.13	-0.047	-0.0058	-0.00071	0.027	0.11	0.0075	0.13	-0.021	-0.036	-0.094	-0.068	-0.026
GrLivArea	-0.0083	0.4	0.26	-0.0096	0.24	0.45	0.57	0.69	0.13	1	0.035	-0.019	0.63	0.42	0.52	0.1	0.83	0.46	0.23	0.47	0.47	0.71
BsmtFullBath	0.0023	0.1	0.16	0.16	-0.42	0.31	0.24	-0.17	-0.047	0.035	1	-0.15	-0.065	-0.031	-0.15	-0.042	-0.053	0.14	0.12	0.13	0.18	0.23
BsmtHalfBath	-0.02	-0.0072	0.048	0.071	-0.096	-0.00031	0.002	-0.024	-0.0058	-0.019	-0.15	1	-0.055	-0.012	0.047	-0.038	-0.024	0.029	-0.077	-0.021	-0.025	-0.017
FullBath	0.0056	0.2	0.13	-0.076	0.29	0.32	0.38	0.42	-0.00071	0.63	-0.065	-0.055	1	0.14	0.36	0.13	0.55	0.24	0.48	0.47	0.41	0.56
HalfBath	0.0068	0.054	0.014	-0.032	-0.041	-0.049	-0.12	0.61	-0.027	0.42	-0.031	-0.012	0.14	1	0.23	-0.068	0.34	0.2	0.2	0.22	0.16	0.28
edroomAbvGr	0.038	0.26	0.12	-0.016	0.17	0.05	0.13	0.5	0.11	0.52	-0.15	0.047	0.36	0.23	1	0.2	0.68	0.11	-0.065	0.086	0.065	0.17
KitchenAbvGr	0.003	-0.0061	-0.018	-0.041	0.03	-0.069	0.068	0.059	0.0075	0.1	-0.042	-0.038	0.13	-0.068	0.2	1	0.26	-0.12	-0.12	-0.051	-0.064	-0.14
TotRmsAbvGrd	0.027	0.35	0.19	-0.035	0.25	0.29	0.41	0.62	0.13	0.83	-0.053	-0.024	0.55	0.34	0.68	0.26	1	0.33	0.15	0.36	0.34	0.53
Fireplaces	-0.02	0.27	0.27	0.047	0.052	0.34	0.41	0.19	-0.021	0.46	0.14	0.029	0.24	0.2	0.11	-0.12	0.33	1	0.047	0.3	0.27	0.47
GarageYrBlt	7.2e-05	0.07	-0.025	-0.088	0.19	0.32	0.23	0.071	-0.036	0.23	0.12	-0.077	0.48	0.2	-0.065	-0.12	0.15	0.047	1	0.59	0.56	0.49
GarageCars	0.017	0.29	0.15	-0.038	0.21	0.43	0.44	0.18	-0.094	0.47	0.13	-0.021	0.47	0.22	0.086	-0.051	0.36	0.3	0.59	1	0.88	0.64
GarageArea	0.018	0.34	0.18	-0.018	0.18	0.49	0.49	0.14	-0.068	0.47	0.18	-0.025	0.41	0.16	0.065	-0.064	0.34	0.27	0.56	0.88	1	0.62
SalePrice	-0.022	0.35	0.26	-0.011	0.21	0.61	0.61	0.32	-0.026	0.71	0.23	-0.017	0.56	0.28	0.17	-0.14	0.53	0.47	0.49	0.64	0.62	1

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:

lambda for ridge and lasso regression during the assignment is as below:



Out[299]:

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	7.086886e-01	7.086881e-01	7.086630e-01
1	R2 Score (Test)	7.359387e-01	7.358763e-01	7.358549e-01
2	RSS (Train)	1.858770e+12	1.858773e+12	1.858933e+12
3	RSS (Test)	7.443129e+11	7.444888e+11	7.445492e+11
4	MSE (Train)	4.266777e+04	4.266781e+04	4.266965e+04
5	MSE (Test)	4.122311e+04	4.122798e+04	4.122965e+04

Analysis:

R2 Score:

➢ The R2 score measures the proportion of the variance in the target variable that is predictable from the predictors.

➢ For both ridge and lasso regression, the R2 scores for both the train and test sets are very close to the R2 score of linear regression. This suggests that ridge and lasso regression have achieved a similar level of predictive performance.

➢ Based on the R2 score alone, there is no significant advantage of using ridge or lasso regression over linear regression.

Residual Sum of Squares (RSS):

➢ The RSS measures the total squared difference between the predicted and actual target values.

- Both ridge and lasso regression have slightly higher RSS values compared to linear regression for both the train and test sets.
- Linear regression has the lowest RSS values, indicating a better fit to the training and test data compared to ridge and lasso regression.

Mean Squared Error (MSE):

- The MSE measures the average squared difference between the predicted and actual target values.
- Similar to the RSS, both ridge and lasso regression have slightly higher MSE values compared to linear regression for both the train and test sets.
- Linear regression has the lowest MSE values, indicating a better fit to the training and test data compared to ridge and lasso regression.

Based on the analysis, linear regression seems to perform slightly better than ridge and lasso regression in terms of the RSS and MSE on both the train and test sets. the trade-off between simplicity and predictive performance. Ridge and lasso regression offer benefits in scenarios where multicollinearity or variable selection is important.

Therefore, As our primary goal is predictive performance, we can choose linear regression. However, As per our concerns about multicollinearity or the interpretability of the model, we can consider ridge regression as it offers regularization to mitigate multicollinearity while maintaining most of the predictors. Lasso regression is also suitable as we suspect that only a subset of predictors is truly important and wish to perform variable selection.

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?

Answer:

The most important predictor variables are as below:
 'TotalBsmtSF','1stFlrSF', 'GarageArea', 'TolRmsAbvGrd' and 'GarageCars'.

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, you can take the following steps:

1. Train/Test Split: we should Split dataset into separate training and testing subsets. The training set is used to train the model, while the testing set is used to evaluate its

performance on unseen data. This helps assess how well the model generalizes to new, unseen observations.

2. Cross-Validation: We have to Utilize techniques like k-fold cross-validation to further assess the model's performance. Cross-validation involves splitting the data into multiple subsets or folds, training the model on different combinations of these folds, and evaluating its performance. This helps estimate the model's performance on unseen data more reliably and reduces the dependency on a single train/test split.
3. Evaluate Multiple Metrics: Use a range of evaluation metrics to assess the model's performance, such as accuracy, precision, recall, F1-score, or mean squared error. Considering multiple metrics provides a comprehensive view of the model's strengths and weaknesses, enabling better understanding of its generalizability.
4. Regularization: We have to apply regularization techniques like ridge regression or lasso regression to control overfitting. Regularization helps prevent the model from fitting the noise in the training data too closely, promoting better generalization to unseen data.
5. Feature Engineering and Selection: We have to engage in proper feature engineering and feature selection techniques to ensure that the model focuses on the most relevant and informative predictors. Removing irrelevant or redundant features can enhance the model's generalizability by reducing noise and overfitting.
6. Test on Diverse Data: We have to assess the model's performance on diverse subsets of the data, ensuring that it performs consistently well across different scenarios. This includes testing the model on data collected at different times, from different sources, or representing different subgroups of the population.

The implications of ensuring model robustness and generalizability for model accuracy are as follows:

1. Lower Over fitting: By focusing on generalization, the model aims to reduce overfitting, which occurs when the model fits the training data too closely and performs poorly on new data. This may result in a decrease in the training set accuracy. However, the model is expected to have better accuracy on unseen data, which is more important for real-world applications.
2. Improved Performance on Unseen Data: By following robust modeling practices, the model's accuracy on unseen data, such as the testing set or new data samples, is expected to be more reliable. This enables the model to make more accurate predictions in real-world scenarios.
3. Reduced Bias: Ensuring that the model is robust and generalizable helps reduce bias in the model's predictions. A bias occurs when the model consistently underestimates or overestimates the target variable. A robust model with good generalization properties is more likely to provide unbiased predictions.

4. Enhanced Model Stability: A robust and generalizable model is less sensitive to small variations in the data. This means that even if the training data changes slightly, the model's performance is expected to be consistent and stable, resulting in improved reliability.

Overall, focusing on model robustness and generalizability may lead to a small reduction in accuracy on the training set but is essential for ensuring better performance and reliability on unseen data, which is crucial for real-world applications and decision-making.