ASSIGNMENT – HOUSE PRICE PREDICTION – PART II

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?

The optimal value of alpha for Ridge and Lasso Regression:

- Ridge - 0.8
- Lasso - 0.000

	Metric	Linear Regression	Lasso Regression	Ridge Regression
0	R2 Score (Train)	0.929	0.935	0.929
1	R2 Score (Test)	0.900	0.906	0.899
2	RSS (Train)	1.388	1.283	1.401
3	RSS (Test)	0.807	0.756	0.814
4	MSE (Train)	0.037	0.001	0.001
5	MSE (Test)	0.044	0.002	0.002

	Metric	Ridge regression	Lasso regression
0	R2 Score (Train)	0.929	0.931
1	R2Score (Test)	0.900	0.905
2	RSS (Train)	1.389	1.364
3	RSS (Test)	0.808	0.771
4	MSE (Train)	0.001	0.001
5	MSE (Test)	0.002	0.002

If we choose double the value of alpha for both ridge and lasso regression:

- Ridge - 0.16
- Lasso - 0.0002

- Effect of doubling alpha in the case of ridge regression-
  - R2 for train remains same whereas R2 for the test has increased
  - RSS has been reduced for both the train and test
  - MSE for both, the train and test remain the same
- Effect of doubling alpha in the case of lasso regression-
  - R2 for both, the train and test remains the same
  - RSS has been increased for both the train and test
  - MSE for both, the train and test remain the same

The most important predictor variables after the change is implemented:

- **OverallQual:** If the Overall quality of the house is Excellent the SalePrice is higher.
- **OverallCond:** If the Overall Condition of house is Excellent the SalePrice is higher.
- **HeatingQC:** Heating quality present in the house has significant increase in the sales price.

	Feature	Coef
19	BsmtFullBath	0.280066
3	OverallCond	0.179230
4	MasVnrArea	0.123003
16	HeatingQC	0.100748
0	MSSubClass	0.073744
2	OverallQual	0.073105
27	GarageQual	0.067138
12	BsmtFinType2	0.053698
56	Neighborhood_Edwards	0.040027
146	SaleType_Oth	0.032514

	Feature	Coefficient
0	LotArea	0.1000
1	OverallQual	0.1928
2	OverallCond	0.1590
3	ExterCond	-0.0280
4	BsmtFinSF1	0.1021
5	BsmtFinSF2	0.0172
6	BsmtUnfSF	0.0165
7	TotalBsmtSF	0.0900
8	HeatingQC	0.0259
9	GrLivArea	0.2822

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?

Since lasso helps in feature reduction as the coefficient value of some of the features becomes zero. Lasso regression provides –

- better R2 score
- lesser RSS

compared to ridge regression. Lasso has a better edge over ridge & should be used as the final model.

	Metric	Linear Regression	Lasso Regression	Ridge Regression
0	R2 Score (Train)	0.929	0.935	0.929
1	R2 Score (Test)	0.900	0.906	0.899
2	RSS (Train)	1.388	1.283	1.401
3	RSS (Test)	0.807	0.756	0.814
4	MSE (Train)	0.037	0.001	0.001
5	MSE (Test)	0.044	0.002	0.002

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?

The five most important predictor variables in the lasso model were removed:

```
lasso_coef.sort_values(by='Coef',ascending=False).head()
```

	Feature	Coef
19	BsmtFullBath	0.280066
3	OverallCond	0.179230
4	MasVnrArea	0.123003
16	HeatingQC	0.100748
0	MSSubClass	0.073744

```
r2_train : 0.9286921628361218
r2_test  : 0.8992204227846299
RSS_train: 1.4013759947886213
RSS_test : 0.8144893078728702
MSE_train: 0.0014141029210783262
MSE_test : 0.0019119467320959395
```

The five most important predictor variables now are:

	Feature	Coef
15	FullBath	0.272400
2	ExterQual	0.191217
0	LotArea	0.097880
13	2ndFlrSF	0.084903
1	OverallQual	0.075442

```
r2_train:  0.925663485001623
r2_test:   0.8971996017590271
RSS_train: 1.4608970317745102
RSS_test:  0.8308213581151602
MSE_train: 0.0014741645123859841
MSE_test:  0.0019502848782046015
```

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?

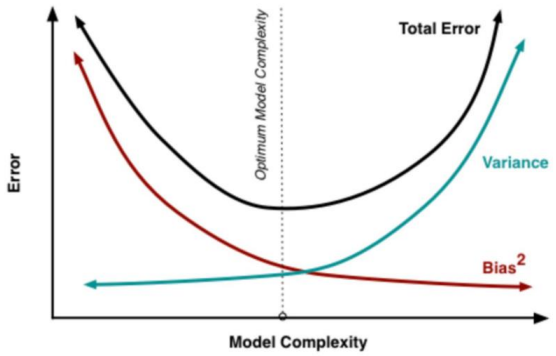
As per Occam's razor- given two models that show similar 'performance' in the finite training or test data, we should pick the one that makes fewer on the test data due to the following reasons—

- Simple models are usually more generic & are more widely applicable
- Simple models require fewer training samples for effective training than the more complex ones & are easier to train.
- Simple model is more robust.
- Complex model tends to change widely with changes in the training data set.
- Simple models have low variance, high bias & complex models have low bias, high variance.
- Simpler models make more errors in the training set. Complex models lead to overfitting as they work very well for training samples, fail measurably when applied to other test samples.

Therefore, to make the model more robust & generalizable, make the model simple but not simpler which will not be of any use.

- Regularization can be used to make the model simpler. Regularization helps to strike the delicate balance between keeping the model simple & not making it too naïve to be of any use. For regression, regularization involves adding a regularization term to the cost that adds up the absolute values or the squares of the parameters of the model.
- Also, making a model simple leads to Bias- Variance trade off, A complete model will need to change for every little change in the dataset & hence is very unstable & extremely sensitive to any changes in the training data. A simpler model that abstracts out some pattern followed by the data points given is unlikely to change widely even if more points are added or removed.
- Bias quantifies how accurate is the model likely to be on test data. A complex model can do an accurate job prediction provided there is enough training data. Models that are too naïve for e.g. One that gives the same answer to all test inputs & makes no discriminations whatsoever has a very large bias as its expected errors across all test inputs are very high.
- Variance refers to the degree of changes in the model itself with respect to changes in the training data.

Thus, the accuracy of the model can be maintaned by keeping the balance between bias & variance as it minimizes the total error as shown in the below graph.



There are different techniques that we need to check while building a robust and generalizable -

- Outlier Treatment
- Missing value treatment
- Significance of predicted variable
- Feature Selection
- Correct algorithm selection
- Cross-validation

Model significance can be determined by P-value, R2, and adjusted R2.

Implications of the accuracy of the model -

- We should have more and more data containing different combinations should that model should learn from different features instead of working with a small dataset.
- Outlier present in the dataset for the feature can affect the model accuracy and should be removed after visualization using boxplot and provide standard data to the model.
- Missing value should be treated based on the type of features and amount of missing value else it leads to an inaccurate model.
- Based on the domain knowledge we need to select the feature which is important and related to the target variable. We can use the VIF value along with the p-value for feature selection. If we can derive more important features from the existing feature which can help us in getting more insight and better relation with the target variable then we should do that.
- Choosing the right algorithm for the problem statement is very crucial for the accuracy of the model. It comes with knowledge and experience.
- More accuracy can sometimes lead to overfitting the model. In those case we use cross validation to get the model with correct accuracy