ASSIGNMENT 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?

Ridge regression optimal value of alpha and its double value:

	params	mean_test_score	mean_train_score	rank_test_score
9	{'ridge__alpha': 5}	0.856129	0.931864	1
10	{'ridge__alpha': 10}	0.856035	0.925625	2

Lasso regression optimal value alpha and its double value:

	params	mean_test_score	mean_train_score	rank_test_score
1	{'lasso__alpha': 0.0005}	0.864214	0.938877	1
2	{'lasso__alpha': 0.001}	0.854128	0.920046	2

Most important predictor variables with optimum alpha and with double of optimum alpha to highlight the changes:

	Variables	Lasso_0005
12	GrLivArea	0.16
64	Neighborhood_Blueste	0.12
2	OverallQual	0.09
295	SaleCondition_AdjLand	0.08
80	Neighborhood_OldTown	0.07
130	RoofMatl_WdShake	0.07
74	Neighborhood_Mitchel	0.06
224	Electrical_FuseF	0.06
187	BsmtCond_Gd	0.06
234	Functional_Maj2	0.06

	Variables	Ridge_05
64	Neighborhood_Blueste	0.12
267	GarageCond_Po	0.11
80	Neighborhood_OldTown	0.09
12	GrLivArea	0.09
2	OverallQual	0.09
74	Neighborhood_Mitchel	0.09
126	RoofMatl_Membran	0.08
283	MiscFeature_Othr	0.08
31	YearRemodAdd	0.08
234	Functional_Maj2	0.06

	Variables	Lasso_001
12	GrLivArea	0.15
2	OverallQual	0.11
64	Neighborhood_Blueste	0.10
295	SaleCondition_AdjLand	0.07
74	Neighborhood_Mitchel	0.07
85	Neighborhood_StoneBr	0.06
130	RoofMatl_WdShake	0.05
187	BsmtCond_Gd	0.05
13	BsmtFullBath	0.05
224	Electrical_FuseF	0.05

	Variables	Ridge_10
64	Neighborhood_Blueste	0.10
2	OverallQual	0.09
12	GrLivArea	0.09
74	Neighborhood_Mitchel	0.08
80	Neighborhood_OldTown	0.07
267	GarageCond_Po	0.06
31	YearRemodAdd	0.06
130	RoofMatl_WdShake	0.05
283	MiscFeature_Othr	0.05
85	Neighborhood_StoneBr	0.05

As depicted in the above images the optimal values of alpha for Ridge is 5 and Lasso is 0.0005.

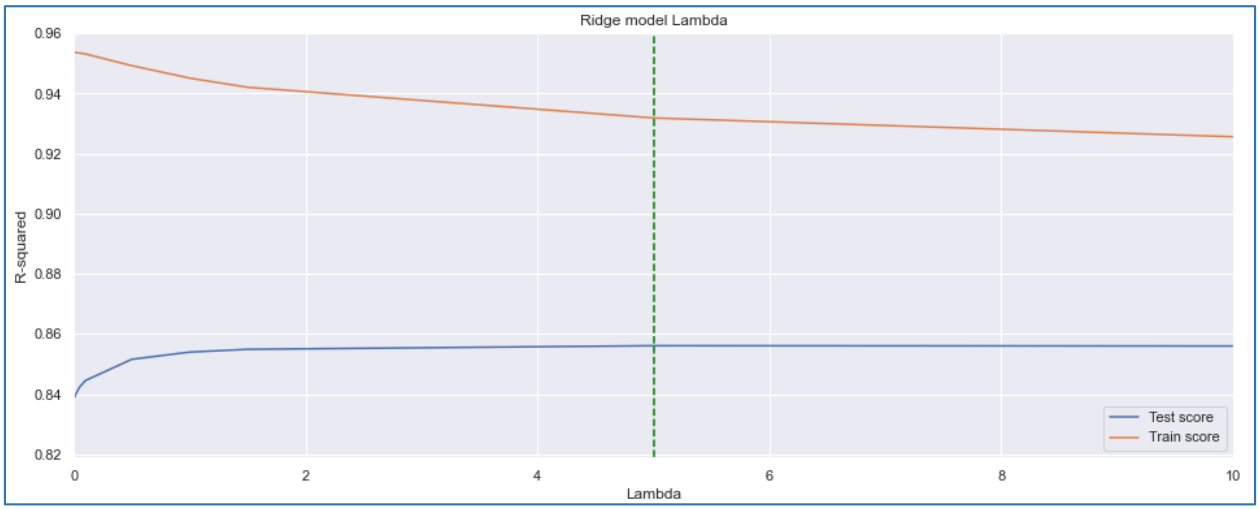
The Mean R2 score with these optimal alphas for train is 0.93 for Ridge and 0.94 for Lasso.

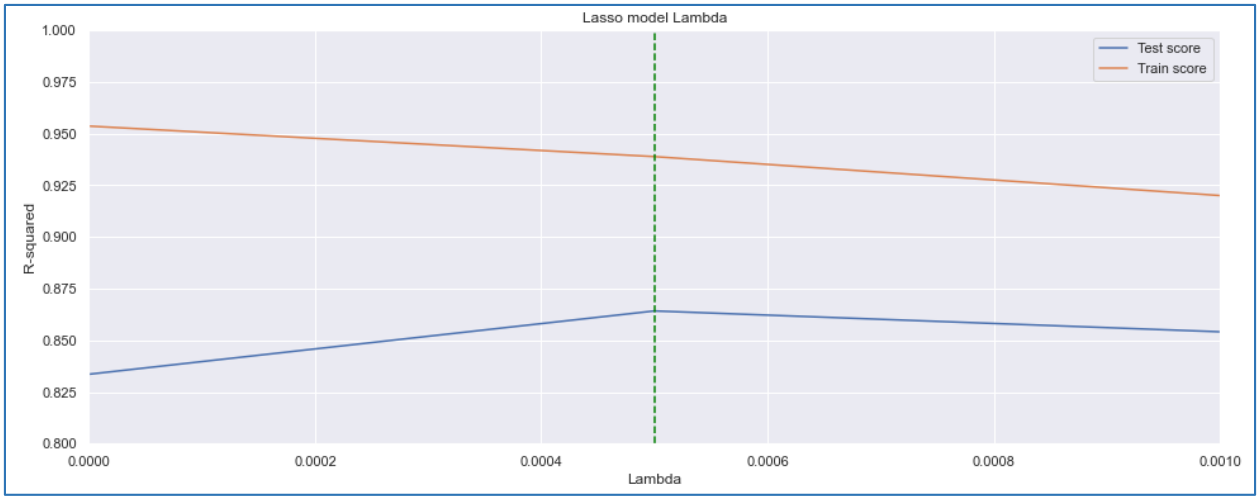
After doubling the values, the alpha values for Ridge and Lasso are 10 and 0.001 respectively.

With this change its observed that R2 score drops a little in case of train data, this gives a clear picture that as the alpha value move towards higher values, shrinkage in penalty term also increases pushing the coefficients towards zero pushing the model towards under fitting.

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?

As shown in below plots the optimal values of lambda for Ridge is **5** & Lasso is **0.0005.**





The Mean R2 score with these optimal lambdas for train is **0.93** for Ridge & **0.94** for Lasso.
It's observed that the accuracy does not affect but the model interpretations become more challenging because of higher number of columns (around **300**) in the data frame after the transformations.
Ridge Regression can shrink the model coefficients close to zero and does not allow feature selection because it retains all the variables present in data. This makes the interpretability a substantial challenge and eventually making difficult to define the business strategies.

Unlike ridge, Lasso regression shrinks the coefficient estimates to zero. Here the penalty term forces some coefficients to zero enabling the feature selection and reduces the multi dimensionality. Hence the models generated with lasso are easy to interpret without affecting the accuracy.

From the business objective its clearly states that the company wants to know the variables that are significant in predicting the price of a house, and how well those variables describe the price of a house.

Hence Lasso regression gives the significant variables that predict the price of house with much easier interpretability. Therefore, Lasso regression inclines more towards the business objective.

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?

As per the business objective lasso model is chosen with the best lambda value as **0.0005**.

With this lambda it's observed that these columns are significant variables that will influence the response variable Sale Price.

	Lasso0005	Variables
12	0.158291	GrLivArea
64	0.115381	Neighborhood_Blueste
2	0.092563	OverallQual
295	0.077167	SaleCondition_AdjLand
80	0.072114	Neighborhood_OldTown

After creating a new model with same lambda value excluding the five important predicted variables earlier. We observe that these columns below are the significant variables that shall influence the response variable Sale Price after the changes.

	coeff	columns
9	0.179468	2ndFlrSF
8	0.140337	1stFlrSF
220	0.089874	Electrical_FuseF
172	0.081374	Foundation_CBlock
279	0.079726	MiscFeature_Othr

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?

If the training and test score is within ~5% then you can say model is generalizable as the model is able to identify the underlying pattern and behave as good as train data on test data as well.
A model Robustness depend on business objective how much accuracy you are looking for, like 80% and above or number of variables in the final model and the coefficient of the Predictors. Moreover, model should be created in such way that it can easily predict un-processed unseen test i.e., test data and shall yield good accuracy. Hence, concept of Pipelines and Column transformers has been used while building robust model.

Implications:

Metrics for Lasso Model with Alpha= 0.0005 gives good implication. As seen the adjusted R2 score does not deviate much from r2 score for train and test which also strengthen the fact that model is robust and more generalized.

	Metric	Linear Regression	Ridge Regression 5	Ridge Regression 10	Lasso Regression 0.0005	Lasso Regression 0.001
0	R2 Score (Train)	0.950333	0.928284	0.922222	0.934403	0.910800
1	R2 Score (Test)	-35056893044.637764	0.887859	0.891552	0.882093	0.888053
2	R2 train/test diff	35056893045.588097	0.040425	0.030670	0.052310	0.022746
3	RSS (Train)	9.065542	13.090082	14.196569	11.973304	16.281490
4	RSS (Test)	1762350402970.436035	5.637454	5.451792	5.927333	5.627694
5	MSE (Train)	0.088100	0.105864	0.110248	0.101248	0.118066
6	MSE (Test)	77688.136710	0.138947	0.136640	0.142475	0.138827
7	Adjusted R2 (Train)	0.946776	0.923147	0.916651	0.929704	0.904411
8	Adjusted R2 (Test)	-47894628526.138916	0.846793	0.851839	0.838915	0.847059