Assignment - Advanced Regression

Submitted By: Gaurav Kalra
Program: Executive PG Programme in Machine Learning & AI - January 2022
Batch: C37
Registered Email: mgkfee@gmail.com
Registered Phone: +91 9739170123

Part-I

Question: Which variables are significant in predicting the price of a house, and How well those variables describe the price of a house.

Total 3 Models were analyzed for the given Problem Statement - Linear, Ridge, and Lasso.

Regression Type	Optimum Value of Alpha	R^2 (R-Squared)	RSS	MSE
Ridge Regression	5.0	0.916290	0.748134	0.041907
Lasso Regression	0.0001	0.920246	0.712783	0.040905
Linear Regression	–	0.722691	2.4783750	0.0058177

After carefully analyzing the performance of all models on Test data, Lasso Model was finalized. And based on Lasso Model, following are the few of top variables that impact the House Pricing.

Variable	Coefficient
GrLivArea	0.248292
TotalBsmtSF	0.082956
YearBuilt	0.076667
OverallQual_3	-0.066269
OverallCond_3	-0.051905
Neighborhood_Crawfor	0.048216
OverallQual_9	0.046631
OverallQual_10	0.042519
GarageArea	0.042169

BsmtFinSF1	0.042113
OverallQual_4	-0.041513

As we notice, that the variables such as GrLivArea, TotalBsmtSF, YearBuilt, Neighborhood_Crawfor, OverallQual_9, OverallQual_10, GarageArea positively impact the SalePrice. Whereas variables such as OverallQual_3 and OverallCond_3 negativvely impact pricing.

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?

Answer-1

Optimal value of alpha for ridge and lasso regression

Regression Type	Optimum Value of Alpha	R^2 (R-Squared)	RSS	MSE
Ridge Regression	5.0	0.916290	0.748134	0.041907
Lasso Regression	0.0001	0.920246	0.712783	0.040905

Both the Regression models were analyzed with following 28 values of alpha using 5-fold k-fold validation technique.

0.0001, 0.001, 0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 2.0, 3.0, 4.0, 5.0, 6.0, 7.0, 8.0, 9.0, 10.0, 20, 50, 100, 500, 1000

Changes in the model with double the value of alpha for both Ridge and Lasso

Regression Type	Double Value of Alpha	R^2 (R-Squared)	RSS	MSE
Ridge Regression	10.0	0.9118656569	0.787679	0.001849
Lasso Regression	0.0002	0.91886814	0.7250959	0.001702

Impact of Doubling alpha on Ridge
Slight decline in R-squared but at the same time RSS has also increased a bit. MSE has come down significantly.

Impact of Doubling alpha on Lasso
Slight decline in R-squared but at the same time RSS has also increased a bit. MSE has come down significantly.

Most important predictor variables after the change is implemented

Ridge with double alpha (5.0)	Lasso with double alpha (0.0002)
GrLivArea	GrLivArea
1stFlrSF	TotalBsmtSF
TotalBsmtSF	OverallQual_9
BsmtFinSF1	YearBuilt
GarageCars	OverallQual_10
2ndFlrSF	OverallQual_3
GarageArea	OverallCond_3
OverallQual_3	GarageArea
TotRmsAbvGrd	Neighborhood_Crawfor
Neighborhood_Crawfor	BsmtFinSF1
OverallQual_9	SaleType_New
OverallQual_4	GarageCars
YearRemodAdd	OverallQual_8
HalfBath	YearRemodAdd
LotArea	OverallQual_4

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

All 3 models were built - Linear Regression, Ridge and Lasso. Following is the table to evaluate the performance of each variable.

Regression Type	Optimum Value of Alpha	R^2 (R-Squared)	RSS	MSE
Ridge Regression	5.0	0.916290	0.748134	0.041907
Lasso Regression	0.0001	0.920246	0.712783	0.040905
Linear Regression	–	0.722692	2.478375	0.076274

After carefully analyzing the results of each model, Lasso seems to be doing the best among all the 3 models tried. Results from Lasso on Test have maximum value of R_2(R-Squared). Also RSS and MSE are minimal in case of Lasso.

For this reason, Lasso will be the preferred model.

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:
After removing the first five important variables (*GrLivArea, TotalBsmtSF, YearBuilt, OverallQual_3, OverallCond_3*) from Lasso Model, the next five most important variables are:

1stFlrSF
2ndFlrSF
BsmtFinSF1
OverallQual_9
OverallQual_10

With these top-5 variables, the new Lasso Model has following results:

R-Squared: 0.9151850908849386
RSS: 0.7580123
MSE: 0.00177937

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:
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?

Following are few of the pointers that may help ensuring a robust and generalizable model:
1. A thorough EDA of the data is a very helpful technique and acts as a foundation for a robust model.
2. Model should account for noise in data, outliers and unavailable values. And such values should be treated using various techniques before building the model.
3. The model must identify correct patterns in the given data.
4. Before applying any regression, pre-conditions should be verified. Such as in case of Linear Regression, the y-variable should be linear. In case non-linearity is present then various techniques such as data transformation should be used. And if non-linearity is still present, then a different approach should be taken such as Polynomial regression or Non-Linear regression.
5. After building the model, cross-validation should be done using techniques such as k-fold validation to ensure that model performs equally well on different combinations of values.
6. After building the models, all assumptions of the Regression must be verified.
7. Model should be simple and not complex.
8. Model must be regularized to avoid the overfitting of the model on train data.
9. Model should show Homoscedastic Unbiased Residuals.
10. For Multiple Linear Regression, with multiple predictors, we should plot the residuals versus the predicted values, the residual plot should show no observable pattern. In case a pattern is observed, it may indicate a problem with some aspect of the linear model.

Implications of a generalizable and robust model on Accuracy and Why

1. It will show low variance on the real-world data.
2. If the error terms are distributed normally then the p-values used to determine coefficients will be reliable and predict accurately.
3. Making sure that there is no non-linearity in data will ensure that predictions by model are accurate.
4. Independence of error terms will ensure that model systematically doesn't underpredicts or overpredicts.
5. Homoscedasticity will ensure that model doesn't give too much weight to a small subset of the data.
6. Normality of Error terms will ensure that model is not built upon large outliers. It will ensure that confidence intervals are not too wide or too narrow.