Assignment part 2

Please note that while building the model, I have updated categorical values by appending the corresponding column name as prefix. So, if the column 'OverallQual' has a category 'Excellent', in the final model it appears to be 'OverallQual Excellent'. I have done this to get more easily understandable values, because many categorical variables have the same kind of values. Say column1 also has a category 'Excellent' along with 3 other columns. By prefixing column name it is easier to understand which particular 'Excellent' we are talking about. Also dummy variable creation won't create variables like 'Excellent_1', 'Excellent_2' etc which were pretty difficult to understand.

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: Through hyperparameter tuning and cross validation the optimal values of alpha for ridge and lasso were 1 and 50 respectively. But as we doubled the values of alpha for both, we got slightly better results. So overall, including hyperparameter tuning and experimentation, the optimal values of alphas were 2 and 100 for ridge and lasso respectively.

Metric	Linear Regression	Ridge Regression alpha 1	Lasso Regression alpha 50	Ridge Regression alpha 2	Lasso Regression alpha 100
R2 Score (Train)	7.170458e-01	9.395560e-01	9.311783e-01	9.323040e-01	9.163797e-01
R2 Score (Test)	5.156019e-01	7.818853e-01	7.517425e-01	7.930183e-01	7.827709e-01

Please note that while listing down the most important predictors, I have selected top 8 important features based on the absolute values of betas. Predictors are listed in descending order of importance.

Ridge regression

Alpha = 1

Important predictors: GrLivArea, Condition2 PosN, OverallQual Very excellent, TotalBsmtSF, GarageArea, LotArea, BsmtFinSF1, OverallQual Excellent [Most important predictor 1st]

Alpha = 2

Important predictors: GrLivArea, Condition2 PosN, OverallQual Very excellent, TotalBsmtSF, GarageArea, BsmtFinSF1, OverallQual Excellent, OverallQual Very good [Most important predictor 1st]

Lasso regression

Alpha = 50

Important predictors: Condition2 PosN, GrLivArea, TotalBsmtSF, GarageArea, LotArea, OverallQual Excellent, BsmtFinSF1, OverallQual Very good [Most important predictor 1st]

Alpha = 100

Important predictors: GrLivArea, Condition2 PosN, TotalBsmtSF, GarageArea, OverallQual Very good, OverallQual Excellent, BsmtFinSF1, Neighborhood Somerst [Most important predictor 1st]

Question2: *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: As explained in the first answer, the optimal values of alphas were 2 and 100 for ridge and lasso respectively.

Metric	Linear Regression	Ridge Regression alpha 1	Lasso Regression alpha 50	Ridge Regression alpha 2	Lasso Regression alpha 100
R2 Score (Train)	7.170458e-01	9.395560e-01	9.311783e-01	9.323040e-01	9.163797e-01
R2 Score (Test)	5.156019e-01	7.818853e-01	7.517425e-01	7.930183e-01	7.827709e-01

I will chose ridge regression with alpha = 2, since the difference between r2 score of train and test is quite low, which implies very low overfitting. Also, the r2 score of train and test is sufficiently high. Based on these two criteria, ridge regression with alpha =2 seems to be the best model.

Though one can argue that lasso regression must have lesser number of predictors since lasso does make coefficients of some of the predictors exactly equal to 0, and thus it should be more generalized model. But the main goal is to check which model explain the variance better, that too on test data. And also, the overfitting is handled quite efficiently in the chosen model.

Question3*: 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: When building the model with lasso, the top 5 important predictor variables were:

Condition2 PosN, GrLivArea, TotalBsmtSF, GarageArea, LotArea and OverallQual Excellent

features	betas
Condition2 PosN	-172548.863948
GrLivArea	164971.476120
TotalBsmtSF	41475.690090
GarageArea	36678.220434
LotArea	29921.089492

After excluding these 5 most important variables and building the lasso model again, we get these 5 new most important predictor variables:

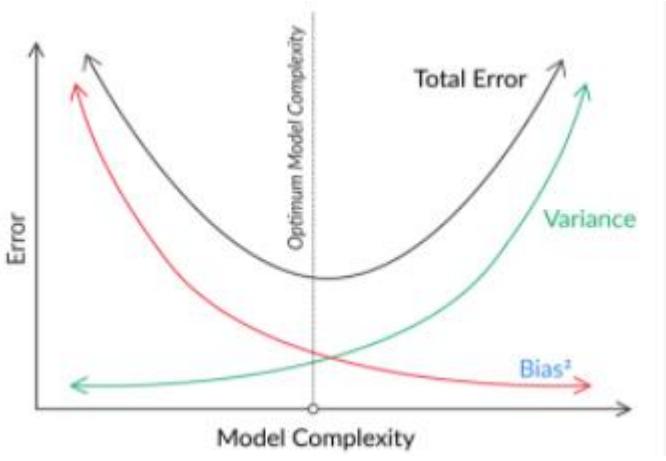
OverallQual Very excellent, BsmtFinSF1, BsmtUnfSF, BsmtFinSF2 and 2ndFlrSF.

features	betas
OverallQual Very excellent	-136389.263422
BsmtFinSF1	102134.476821
BsmtUnfSF	64772.073034
BsmtFinSF2	60691.216074
2ndFlrSF	55738.244854

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: We can make a model more robust and generalisable by creating a model by not having too many predictors and keeping the coefficients of those predictors as small as possible. Why? In that case even if the training dataset changes a little bit, it won't impact the model a lot. The accuracy may suffer a little bit but it is acceptable to compromise a little bit of accuracy to make model more generalizable.

And that's the trade off we make in general. We reduce the variance a little bit, in turn our model gain a little bit of bias. All this to make our model more general and robust.



This image is taken from one of the upgrad lectures. When we move from right to left, the model complexity decreases, and that leads to increase in bias but reduction in variance. We chose a point where the total error is minimum. Increase in bias leads to little bit of compromise in model accuracy.