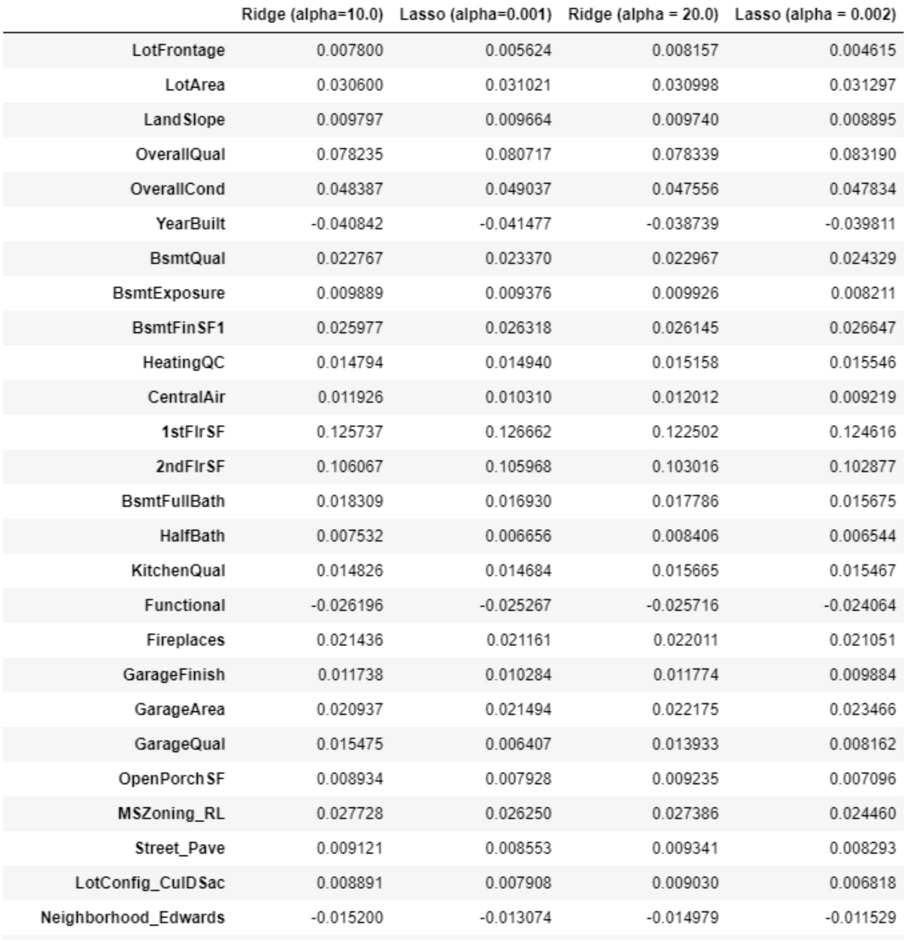
What is the optimal value of alpha for ridge and lasso regression? What will be the changes in themodel if you choose to double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?

Ans.

The optimal value of alpha for Ridge is 10 and for Lasso it is 0.001. With these alphas the R2 of the model was approximately 0.87. After doubling the alpha values in the Ridge and Lasso, the predictionaccuracy remains around 0.87 but there is a small change in the co-efficient values. Below are the changes in the co-efficient



	Ridge (alpha=10.0)	Lasso (alpha=0.001)	Ridge (alpha = 20.0)	Lasso (alpha = 0.002)
LotFrontage	0.007800	0.005624	0.008157	0.004615
LotArea	0.030600	0.031021	0.030998	0.031297
LandSlope	0.009797	0.009664	0.009740	0.008895
OverallQual	0.078235	0.080717	0.078339	0.083190
OverallCond	0.048387	0.049037	0.047556	0.047834
YearBuilt	-0.040842	-0.041477	-0.038739	-0.039811
BsmtQual	0.022767	0.023370	0.022967	0.024329
BsmtExposure	0.009889	0.009376	0.009926	0.008211
BsmtFinSF1	0.025977	0.026318	0.026145	0.026647
HeatingQC	0.014794	0.014940	0.015158	0.015546
CentralAir	0.011926	0.010310	0.012012	0.009219
1stFlrSF	0.125737	0.126662	0.122502	0.124616
2ndFlrSF	0.106067	0.105968	0.103016	0.102877
BsmtFullBath	0.018309	0.016930	0.017786	0.015675
HalfBath	0.007532	0.006656	0.008406	0.006544
KitchenQual	0.014826	0.014684	0.015665	0.015467
Functional	-0.026196	-0.025267	-0.025716	-0.024064
Fireplaces	0.021436	0.021161	0.022011	0.021051
GarageFinish	0.011738	0.010284	0.011774	0.009884
GarageArea	0.020937	0.021494	0.022175	0.023466
GarageQual	0.015475	0.006407	0.013933	0.008162
OpenPorchSF	0.008934	0.007928	0.009235	0.007096
MSZoning_RL	0.027728	0.026250	0.027386	0.024460
Street_Pave	0.009121	0.008553	0.009341	0.008293
LotConfig_CulDSac	0.008891	0.007908	0.009030	0.006818
Neighborhood_Edwards	-0.015200	-0.013074	-0.014979	-0.011529

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 to and why?

Ans.

The optimum lambda value in case of Ridge and Lasso is as

below.Ridge – 10

Lasso – 0.001

The Mean Squared Error in case of Ridge and Lasso are.

Ridge - 0.8704201226046137

Lasso - 0.8745163217343286

Lasso Regression produced slightly higher R2 score on test data than Ridge Regression. Choosing Lasso as the final model.

Question 3:

After building the model, you realised that the five most important predictor variables in the lassomodel are not available in the incoming data. You will now have to create another model excludingthe five most important predictor variables. Which are the five most important predictor variablesnow?

Ans.

The five most important predictor variables in the current lasso model are as below.

'1stFlrSF'

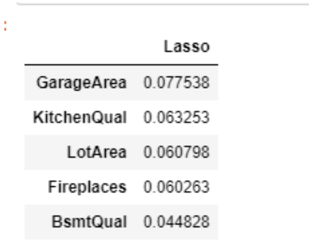
'2ndFlrSF'

'OverallQual'

'OverallCond',

'SaleCondition_Partial'

The new Top 5 predictors are as below.



	Lasso
GarageArea	0.077538
KitchenQual	0.063253
LotArea	0.060798
Fireplaces	0.060263
BsmtQual	0.044828

Question 4

How can you make sure that a model is robust and generalisable? What are the implications of thesame for the accuracy of the model and why?

The mode is robust and generalizable when.

1. Test accuracy is not much lesser than the training score

2. The model should not be impacted by the outliers: Outlier treatment is most important to get the robust model. We can detect outliers in the dataset using box plots, Z score etc.

Treating the outliers will not affect mean, median etc. so that we can impute correct values to missing values., the outlier analysis needs to be done and only those which are relevant. This would help standardize the predictions made by the model. If the model is not robust, it cannot be trusted for predictive analysis.

3. The predicted variables should be significant. Model significance can be determined by the P-values, R2 and adjusted R2. Always a simple model can be more robust.

Implications of Accuracy of a model:

1. **Gain as much data as much you can**: Having more data allows the data to train itself, instead of depending on the weak correlations and assumption, it is good to have more data.

2. **Fix missing values and outliers**: If the data has missing values and outliers can lead to inaccurate models. Outliers can affect the mean, median that we are imputing to continuous variables. You can get the outlier values using a boxplot, treating the outliers in the data will make our mode more accurate.

3. **Featuring Engineering or newly derived columns/Standardize the values**: We can extract the new data from the existing data ex: from DOB we can get the Age of the person, after extracting the new data required, we can drop the existing features. Scaling the values: ex: one value is in meters, the other is Kilo meters, it is important to scale these features into one standardized unit. If we did this, we could get an accurate model.

4. **Feature Selection:** It is purely based on domain knowledge, so that we can select important features that have a good impact on the target variable. Data visualization also helps the selecting the features. Statistical parameters like p-Values and VIF can give us significant variables.

5. **Applying the right algorithm**: Choosing the right machine learning algorithm is very important to get an accurate model. This will come with experience.

6. **Cross validation:** Sometimes more accuracy will cause overfitting, then we can use cross validation technique, i.e. leave a sample on which you do not train the model & test the model on this sample before got to the final model