Assignment: Advanced Regression- Subjective Questions

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 & Explanation

The optimal value of **Alpha** for,

- **Ridge** regression:
  - **20**, when all features are scaled including target SalePrice scaled.
  - **20**, when all features are scaled except target SalePrice.
- **Lasso** regression
  - **0.001**, when all features are scaled including target SalePrice scaled.
  - **100**, when all features are scaled except target SalePrice.

The alpha values are high. Higher alpha increases the effect of regularization.

After doubling Alpha values for both Ridge and Lasso's RSS and RMSE values are increased. However, R2 Score is similar.

The RMSE value tells us that the average deviation between the predicted house price made by the model and the actual house price. RMSE is high indicates there is huge deviation between predicted and actual house prices.

The R2 value tells us that the predictor variables in the model (Neighboorhood, GrLivArea, OverallQual, SaleType and more) can explain 86+% of the variation in the house prices.

The most important predictor variables are,

Neighborhood, GrLiveArea, SaleType, OverallQual, Condition1, ...

	Linear	Ridge	Lasso
cat__RoofMatl_WdShngl	6.376605e+08	0.129257	0.731956
cat__Neighborhood_StoneBr	1.449152e+09	0.185928	0.590677
cat__Neighborhood_NridgHt	1.449152e+09	0.193369	0.441871
cat__Neighborhood_NoRidge	1.449152e+09	0.183592	0.419142
num__GrLivArea	-5.925993e+10	0.158532	0.313263
cat__SaleType_New	3.174911e+09	0.084357	0.243421
cat__Neighborhood_Crawfor	1.449152e+09	0.113117	0.227147
num__OverallQual	1.327372e-01	0.193173	0.165290
cat__Exterior1st_BrkFace	3.177239e+09	0.099380	0.152677
cat__BldgType_2fmCon	1.746807e+10	0.066676	0.141912
cat__Condition1_Norm	4.050167e+09	0.110420	0.115944
cat__Neighborhood_Somerst	1.449152e+09	0.001417	0.091042
cat__LotConfig_CulDSac	-3.904514e+09	0.077034	0.088778
cat__Neighborhood_BrkSide	1.449152e+09	0.055016	0.083690
num__BsmtExposure	7.186127e-02	0.081386	0.083035

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

Multivariate linear regression indicates signs of multicollinearity – i.e. many predictor variables are highly correlated to each other, such that they do not provide unique or independent information in the regression model.

This is causing the coefficient estimates of the model to be unreliable and have high variance. When the model is applied on unseen data, it is performing poorly.

After applying ridge regression and lasso regression.

- Ridge regression: RSS + $\lambda\Sigma\beta_j^2$

- Lasso regression: RSS + $\lambda\Sigma|\beta_j|$

The shrinkage penalty (λ) approaches infinity, the shrinkage penalty becomes more influential and the predictor variables that aren't important in the model shrink towards zero.

In Lasso regression, it's possible that some of the coefficients could go completely to zero when λ gets sufficiently large.

Ridge vs. Lasso Regression

After performing k-fold cross-validation, we have to choose model that produces the lowest test mean squared error.

In our case, **Lasso Regression has lowest MSE error.**

	Metric	Linear Regression	Ridge Regression	Lasso Regression
0	R2 Score (Train)	0.921	0.874	0.911
1	R2 Score (Test)	-636705653632319488.000	0.887	0.898
2	RSS (Train)	92.092	146.621	103.916
3	RSS (Test)	18764977179823452 9792.000	33.166	30.065
4	MSE (Train)	0.281	0.355	0.299
5	MSE (Test)	801645934.804	0.337	0.321

Lasso regression tends to perform better only when a small number of predictor variables are significant, because it's able to shrink insignificant variables completely to zero and remove them from the model.

However, when many predictor variables are significant in the model and their coefficients are roughly equal then ridge regression tends to perform better because it keeps all of the predictors in the 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 & Explanation

Based on “model.coef_” values, below features are significant after first 5 features,

1. LotConfig
2. YearBuilt
3. MasVnrArea
4. BsmtFinSF1
5. BsmtExposure

cat__LotConfig_CulDSac	-3.904514e+09	0.097713	0.111421
num__YearBuilt	1.359723e-01	0.058869	0.104866
cat__MasVnrType_None	-1.266124e+10	0.075677	0.104042
num__MasVnrArea	8.530399e-02	0.083312	0.085570
num__BsmtFinSF1	7.989667e+10	0.019325	0.085363
num__BsmtExposure	7.186127e-02	0.083899	0.078632
cat__Neighborhood_Somerst	1.449152e+09	-0.000851	0.075555
cat__Exterior1st_CemntBd	3.177239e+09	0.061981	0.073823
cat__GarageType_NA	1.493932e+10	0.049427	0.071488
num__OverallCond	7.461929e-02	0.054900	0.069864
cat__SaleCondition_Normal	-6.613529e+09	0.018632	0.062459
num__GarageArea	6.479500e-02	0.064468	0.060938
cat__LotShape_IR2	-2.896926e+10	0.089767	0.058299
num__TotRmsAbvGrd	4.539730e-02	0.080305	0.056858

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

Multivariate linear regression model with too many correlated features suffers multicollinearity. As a result, the model performs good on training data and performs bad on test dataset.

Multicollinearity causes the coefficient estimates of the model to be unreliable and have high variance.

To avoid overfitting to train dataset, use Lasso regression and Ridge regression regularization methods to constrain or regularize the coefficient estimates of certain predictor variables or features of the model.

Both these models try to minimize the sum of squared residuals (RSS) by using penalty term. By using k-fold cross-validation to choose the model that produces the lowest test mean squared error.

As a result, the accuracy may have negatively impacted but balances both bias and variance by changing model coefficients.