Upgrad Assignment – Regularized Linear Regression Subjective Questions

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?

Ans:
Ridge Optimal Alpha : 3.0

Lasso Optimal Alpha: 0.001

On doubling the alpha, the models perform slightly worse on the training set and test set.

Ridge: The coefficients of the features / model shrink further and get smaller.

Lasso: More feature coefficients are assigned zero coefficients.
Optimal Alpha: **129** have zero coefficients out of 189 features.
Double Alpha: **146** have zero coefficients out of 189 features.

Feature Importance After the change has been made:

Ridge:

	Feature	Coef	abs_coeff
23	1stFlrSF	0.276605	0.276605
6	OverallQual	0.249783	0.249783
25	GrLivArea	0.242000	0.242000
7	OverallCond	0.162258	0.162258
17	BsmtFinSF1	0.156875	0.156875
35	GarageArea	0.141116	0.141116
27	FullBath	0.135164	0.135164
24	2ndFlrSF	0.133616	0.133616
1	LotArea	0.127749	0.127749
72	Neighborhood_Crawfor	0.110530	0.110530

Lasso:

	Feature	Coef	abs_coeff
25	GrLivArea	0.613745	0.613745
6	OverallQual	0.496547	0.496547
23	1stFlrSF	0.258086	0.258086
35	GarageArea	0.180660	0.180660
7	OverallCond	0.127521	0.127521
17	BsmtFinSF1	0.124733	0.124733
1	LotArea	0.092415	0.092415
72	Neighborhood_Crawfor	0.086593	0.086593
13	BsmtQual	0.085968	0.085968
181	SaleType_New	0.077747	0.077747

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: I'll choose Lasso, as it has the property to drive model coefficients to zero, and enables feature selection and reducing model complexity.
Metrics: Training R-Squared:0.923, Test R-Squared:0.922
Train RMSLE:0.108 Test RMSLE:0.117

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:
Top 5 Features after removing the most important features:
'BsmtFinSF1', 'BsmtUnfSF', '2ndFlrSF', 'FullBath', 'LotArea'.

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:

The key factor in making sure that the model is robust and generalizable is to get the bias variance trade-off right, and make sure the model doesn't underfit or overfit.

Regularization and cross-validation can help to prevent overfitting by encouraging the model to use all the input features in a more balanced way and by evaluating its performance on new data.

The implications of these techniques for the accuracy of the model are that they can help to prevent overfitting or underfitting, which can lead to poor performance on new, unseen data. Overfitting occurs when a model is too complex and fits the training data too closely, resulting in poor performance on new data. Underfitting occurs when the model fails to learn any data patterns.