Acknowledgement: Your acknowledgement for collaboration and other sources goes here.

Name: Badri Narayanan

Advanced Regression - Assignment 2

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?

The optimal value of alpha for both Ridge and Lasso regression is 0.0001 which was obtained through hyper-parameter tuning, where GridSearchCV was used to find the best value of alpha.
When we double the value of alpha in both the cases, the r2 scores and RSS values remain more or less the same. There is a difference in the coefficient values however and the values are listed below.

	Linear	Ridge	Ridge_2xalpha	Ridge_diff	Lasso	Lasso_2xalpha	Lasso_diff
OverallQual_10	1.037463	1.037320	1.037178	-1.426756e-04	0.776095	0.630553	-0.145542
OverallQual_9	1.009201	1.009073	1.008945	-1.279374e-04	0.768515	0.642886	-0.125629
OverallQual_8	0.814445	0.814321	0.814196	-1.247153e-04	0.578182	0.456509	-0.121673
OverallQual_7	0.688270	0.688147	0.688024	-1.228428e-04	0.454339	0.334481	-0.119858
OverallQual_6	0.567621	0.567499	0.567377	-1.217587e-04	0.334316	0.214738	-0.119578
OverallQual_5	0.474533	0.474412	0.474291	-1.208158e-04	0.242209	0.123260	-0.118949
RoofMatl_Membran	0.427149	0.427086	0.427023	-6.308057e-05	0.225637	0.027608	-0.198029
Heating_Wall	0.346692	0.346654	0.346616	-3.766955e-05	0.176469	0.000000	-0.176469
GrLivArea	0.149446	0.149447	0.149449	1.276386e-06	0.151113	0.152189	0.001076
OverallQual_4	0.362426	0.362307	0.362188	-1.190032e-04	0.132975	0.016649	-0.116326
MSZoning_FV	0.089389	0.089390	0.089391	9.901537e-07	0.087400	0.085156	-0.002244
Neighborhood_Crawfor	0.088949	0.088948	0.088948	-3.453348e-07	0.084948	0.081372	-0.003575
TotalBsmtSF	0.082249	0.082251	0.082252	1.382673e-06	0.084226	0.085801	0.001574
MSSubClass_SPLIT FOYER	0.078804	0.078804	0.078805	8.611076e-07	0.074535	0.069684	-0.004851
OverallQual_3	0.299655	0.299536	0.299418	-1.184004e-04	0.060413	-0.022411	-0.082824
GarageQual_Fa	-0.084591	-0.084591	-0.084590	3.948483e-07	-0.081663	-0.076989	0.004673
MSZoning_RH	-0.099264	-0.099261	-0.099258	2.963910e-06	-0.083706	-0.072699	0.011006
GarageCond_Fa	-0.094755	-0.094752	-0.094750	2.414711e-06	-0.085292	-0.084224	0.001068
MSSubClass_2-1/2 STORY ALL AGES	-0.106143	-0.106140	-0.106137	2.969251e-06	-0.088061	-0.068946	0.019115
MSSubClass_2-STORY PUD-1946 & NEWER	-0.102438	-0.102435	-0.102433	2.223420e-06	-0.094811	-0.087865	0.006946
MSZoning_RM	-0.109455	-0.109455	-0.109456	-2.237234e-07	-0.111640	-0.114163	-0.002523
OverallCond_4	-0.140631	-0.140631	-0.140632	-6.895972e-07	-0.139150	-0.139108	0.000043
Functional_Mod	-0.216073	-0.216052	-0.216032	2.069997e-05	-0.142833	-0.073293	0.069539
Electrical_FuseP	-0.339816	-0.339780	-0.339743	3.642009e-05	-0.205069	-0.070913	0.134156
OverallCond_3	-0.263898	-0.263915	-0.263932	-1.690690e-05	-0.293457	-0.311375	-0.017919
Heating_Grav	-0.289626	-0.289635	-0.289644	-8.990551e-06	-0.295308	-0.293205	0.002103
Functional_Maj2	-0.372647	-0.372633	-0.372619	1.415079e-05	-0.328473	-0.297885	0.030588
Functional_Sev	-0.525499	-0.525449	-0.525398	5.041183e-05	-0.399010	-0.271554	0.127456
Condition2_PosN	-0.890429	-0.890373	-0.890318	5.533374e-05	-0.811210	-0.730365	0.080845

We can observe that in Lasso regression, the positive coefficients values have decreased except for these two variables - GrLivArea and TotalBsmtSF. The beta value for the variable Heating_Wall has been pushed to 0 as well. This just means that as alpha is increased, more the regularization, and hence the coefficients are penalised. The most important predictor

variables seem to be the OverallQual (Overall Quality of the house), followed by RoofMaterial_Membran (Membrane being used as the roof material) and GrLivArea (Greater Living area in sq.feet).

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?

As stated earlier, we have obtained lambda value as 0.0001 for both Ridge and Lasso regression. Lasso regression allows us to perform feature elimination as the coefficients are pushed to 0 with increase in lambda, i.e. more regularization, which allows us to get away with making the model more complex. In our case, we can choose to go with lasso regression for the above stated reason but in general, we need to consider the domain and data that we are dealing with and choose the type and value of regularization more carefully.

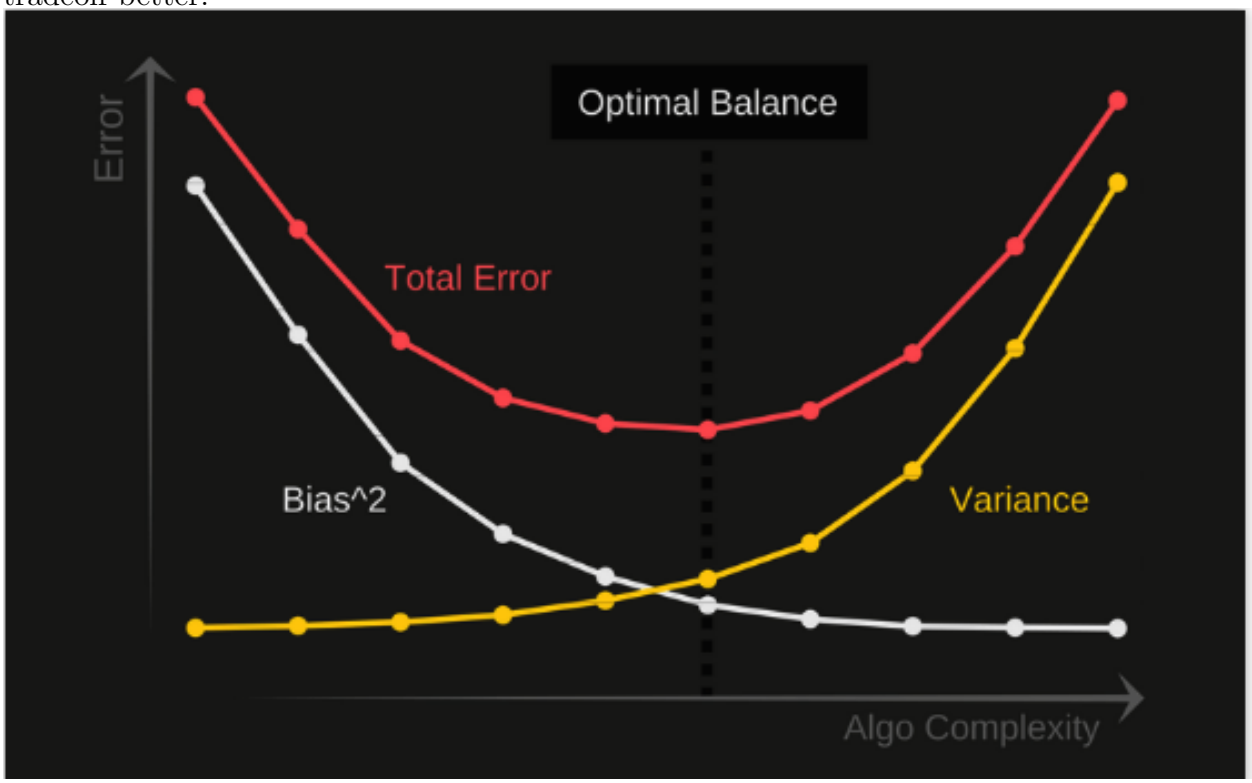
After building the model, you realised that the five most important predictor variables in the lasso model are not available in the incoming data. You will now have to create another model excluding the five most important predictor variables. Which are the five most important predictor variables now?

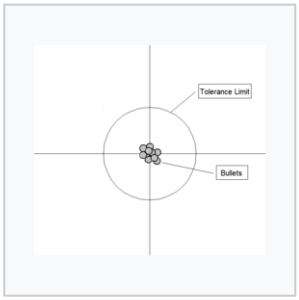
TODO

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?

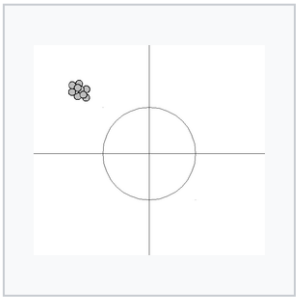
In order to make a more generalized and robust model, we have to understand the bias-variance trade off of each model that we're building. In short, a more complex model is going to have high bias and less variance and a simpler model is going to have low bias and high variance.
What this means is that simpler models make more errors in the training set, but the predicted values for an unknown test set might not be all that bad. However, complex models tend to overfit, which means that they work very well for the training set and we could achieve practically near-zero error, but when it comes to prediction on an unknown distribution, the model is going to do very poorly. Techniques like regularization help us develop more complex models, however we need to choose carefully on how complex we let our model be and how much we penalize it. Training accuracy could be really high for an over fitted and complex model but the testing accuracy would be low. For simpler and general models, we are going to have high training error as stated above, but testing accuracy could

end up being lesser than complex models. Below is a diagram to understand bias-variance tradeoff better.

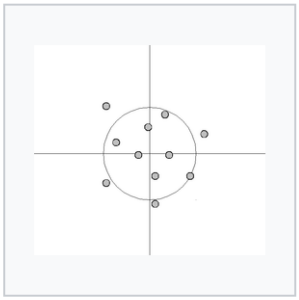




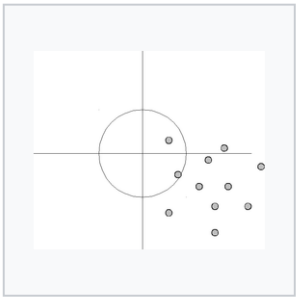
bias low, variance low



bias high, variance low



bias low, variance high



bias high, variance high