Problem Statement - Part II

Assignment Part-II

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?

Ans: Optimal value of alpha for Ridge is 9.0 and for Lasson Regression is 0.0001.

If values of alpha for both ridge and lasso are doubled, the model coefficients are altered.

	Ridge (alpha=9.0)	Lasso (alpha=0.0001)	Ridge (alpha = 18.0)	Lasso (alpha = 0.0002)
MSSubClass	-0.006491	-0.007315	-0.005318	-0.006728
LotArea	0.034155	0.033385	0.034844	0.033426
LandSlope	0.008769	0.008699	0.008819	0.008702
OverallQual	0.078208	0.078486	0.078148	0.078820
OverallCond	0.050032	0.050763	0.049267	0.050701
YearBuilt	-0.039039	-0.040784	-0.037541	-0.040809
BsmtQual	0.019073	0.019358	0.019186	0.019716
BsmtExposure	0.009381	0.009327	0.009374	0.009273
BsmtFinSF1	0.033459	0.032500	0.033897	0.032131
BsmtUnfSF	0.009571	0.008196	0.010200	0.007574
HeatingQC	0.014708	0.014429	0.015041	0.014537
CentralAir	0.010831	0.010492	0.011016	0.010343
1stFlrSF	0.121553	0.124911	0.118598	0.124983
2ndFlrSF	0.108053	0.110863	0.105040	0.110329
BsmtFullBath	0.019707	0.019837	0.019291	0.019482
HalfBath	0.008273	0.007387	0.009030	0.007334
KitchenQual	0.014386	0.013625	0.015182	0.013730
Functional	-0.025925	-0.026367	-0.025463	-0.026341
Fireplaces	0.020850	0.020241	0.021480	0.020322
GarageFinish	0.010554	0.009994	0.011039	0.009986
GarageArea	0.022149	0.021113	0.023189	0.021270
GarageQual	0.017915	0.017649	0.016473	0.015529
OpenPorchSF	0.008305	0.007873	0.008683	0.007836
MSZoning_RL	0.028272	0.028274	0.028108	0.028099
Street_Pave	0.009031	0.008864	0.009135	0.008810
LotConfig_CulDSac	0.006914	0.006749	0.007017	0.006710
Neighborhood_Edwards	-0.016154	-0.016099	-0.015925	-0.015751
Neighborhood_NAmes	-0.010657	-0.010414	-0.010568	-0.010063
Neighborhood_NWAmes	-0.006727	-0.006835	-0.006403	-0.006540
Neighborhood_NridgHt	0.014832	0.014906	0.014640	0.014757
Neighborhood_Somerst	0.024268	0.024516	0.023835	0.024298
Condition1_Feedr	0.010865	0.010964	0.010534	0.010710
Condition1_Norm	0.024093	0.024372	0.023584	0.024121
Condition2_Norm	0.008973	0.008816	0.009019	0.008723
BldgType_TwnhsE	0.006737	0.007137	0.005952	0.006632
RoofStyle_Gable	-0.021744	-0.022714	-0.019554	-0.020749
RoofStyle_Hip	-0.016818	-0.018111	-0.014299	-0.016122
Exterior1st_HdBoard	-0.017192	-0.017293	-0.016015	-0.015900
Exterior1st_Wd Sdng	-0.018244	-0.018044	-0.017696	-0.017288
Exterior2nd_HdBoard	0.009517	0.009713	0.008251	0.008314
Exterior2nd_Wd Sdng	0.013516	0.013424	0.012761	0.012501
MasVnrType_BrkFace	0.015279	0.015835	0.011951	0.010473
MasVnrType_None	0.015343	0.016218	0.011624	0.010668
MasVnrType_Stone	0.012189	0.012461	0.010177	0.009089
Foundation_PConc	0.018036	0.017690	0.018472	0.017884
Heating_GasA	-0.008899	-0.008620	-0.008964	-0.008423
GarageType_Not_applicable	0.008014	0.007267	0.006967	0.005169
PavedDrive_Y	0.010204	0.009890	0.010330	0.009726
SaleCondition_Normal	0.029489	0.029970	0.028907	0.029885
SaleCondition_Partial	0.034436	0.034925	0.033854	0.034843

Table 1. Coeff comparison with various Lasso and Ridge values

Most important predictor variable after change is implemented is “1stFlrSF”. Its coefficient is most significant before as well as after the change is implemented.

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?

Ans:

With Ridge Regression: 0.871704

With Lasso Regression: R2 score (test) : 0.871732

Lasso regression chosen during the assignment because it has higher R^2 on the test-data.

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?

Ans: Ref to table 1. Most important predictors with lasso reg with (alpha=0.0001): 1stFlrSF, 2ndFlrSF, OverallQual, OverallCond, SaleCondition_Partial.

If these predictors are not available (dropped) and need to select another predictor. After creating training data without above predictors, new five most important predictors are: BsmtFinSF1, LotArea, BsmtUnfSF, GarageArea, KitchenQual.

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?

Ans: Generalisable and robustness imply that the model is not Overfitting or it is not too complex. This can be achieved with high bias and low variance, which would result in less dynamic performance of the model. However, over-doing this would result in the “over-simplification” of the model and the model might not be able to capture all the important patterns form the training data. The implication of this is that the accuracy of the model is very poor as the model is not adequately trained.

Regularization helps to achieve a sweet spot where there is a “appropriate” trade-off between the Bias and variance so that the overall error is minimum without making the model overly complex.