Advanced Regression Assignment Part – II

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: Optimal Values for Alpha:
   a) LASSO: .001
   b) Ridge: 6.0

   When you double up the above Ridge Alpha value to 12: The important predictor terms that emerged out are:


```
{'LotFrontage': -0.003174703981713132, 'LotArea': 0.0012925222884966699, 'OverallQual':
0.08018363830600277, 'OverallCond': 0.027139004474747232, 'YearBuilt': 0.001782027795365031,
'YearRemodAdd': 0.028311465137897646, 'MasVnrArea': 0.006912895231873053, 'ExterQual':
0.019414526897083888, 'ExterCond': -0.014686947334969202, 'BsmtQual': 0.036141397351883 53,
'BsmtCond': 0.009710944889490298, 'BsmtFinType1': 0.032409604418151654, 'BsmtFinSF1':
0.006768940716745039, 'BsmtFinType2': 0.0014421258491268197, 'BsmtFinSF2': -
0.002491032386287739, 'BsmtUnfSF': 0.014481339082574499, 'TotalBsmtSF': 0.010059397423010874,
'HeatingQC': 0.0041849855666114474, '1stFlrSF': 0.026575587044775747, '2ndFlrSF':
0.035833868330018616, 'LowQualFinSF': -0.0037827367972067464, 'GrLivArea':
0.034861397520119144, 'KitchenQual': 0.030403864518487485, 'FireplaceQu':
0.042609798473402995, 'GarageYrBlt': 0.014765889136271266, 'GarageArea': 0.035613741389045335,
'GarageQual': 0.012459084283630318, 'GarageCond': 0.011822374920680024, 'PavedDrive':
0.006815869266898619, 'WoodDeckSF': 0.025336746632662346, 'OpenPorchSF':
0.0040828138656003295, 'EnclosedPorch': -0.015633350177225804, '3SsnPorch':
0.008689625004487323, 'ScreenPorch': 0.01733922128804006, 'PoolArea': -0.018677817728019228,
'PoolQC': -0.018677817728019203, 'Fence': -0.02160961535671231, 'MiscVal':
0.0017208205069603744, 'MoSold': 0.008236181058835973, 'YrSold': -0.0047986879091725726,
'MSSubClass_120': 0.008241484642363543, 'MSSubClass_160': -0.02762028038784328,
'MSSubClass_180': -0.01058514209505515, 'MSSubClass_190': -0.0034403968815518 72,
'MSSubClass_20': 0.011934825311001256, 'MSSubClass_30': -0.025269729111057095,
'MSSubClass_40': 0.007299854859862752, 'MSSubClass_45': 0.005608323028445493, 'MSSubClass_50':
0.0020814379865327676, 'MSSubClass_60': 0.00067186042073621 77, 'MSSubClass_70':
0.029913881481539714, 'MSSubClass_75': -0.0013804099665590793, 'MSSubClass_80': -
0.0033313173457887256, 'MSSubClass_85': 8.804190748797212e-05, 'MSSubClass_90':
0.005787566149885229, 'MSZoning_C (all)': 0.0, 'MSZoning_FV': 0.013485785664331843,
'MSZoning_RH': -0.008385574656911896, 'MSZoning_RL': 0.013220912436484773, 'MSZoning_RM': -
0.018321123443904446, 'Street_Grvl': 0.002501281351078212, 'Street_Pave': -
0.002501281351078204, 'Alley_Grvl': 0.006304474497165772, 'Alley_Pave': -0.004138308055653146,
'Alley_nan': -0.002166166441512555, 'LotShape_IR1': 0.006239110923489184, 'LotShape_IR2':
0.004490193699867 19, 'LotShape_IR3': -0.019013210012869087, 'LotShape_Reg':
0.008283905389513431, 'LandContour_Bnk': -0.024601870068470644, 'LandContour_HLS':
0.002106537791888006, 'LandContour_Low': 0.010450751515481567, 'LandContour_Lvl':
0.012044580761101261, 'Utilities_AllPub': 0.0, 'Utilities_NoSeWa': 0.0, 'LotConfig_Corner': -
0.0067961088126855035, 'LotConfig_CulDSac': 0.021245639045026552, 'LotConfig_FR2': -
0.008105258435391617, 'LotConfig_FR3': -0.005862950068841757, 'LotConfig_Inside': -
0.00048132172810840256, 'LandSlope_Gtl': -0.02333288998957003, 'LandSlope_Mod':
0.03499866901496902, 'LandSlope_Sev': -0.011665779025399071, 'Neighborhood_Blmngtn': -
0.013826758528755945, 'Neighborhood_Blueste': 0.0, 'Neighborhood_BrDale':
0.004078579568495988, 'Neighborhood_BrkSide': 0.006610230625830612, 'Neighborhood_ClearCr':
0.026859505205359988, 'Neighborhood_CollgCr': -0.005031792428155305, 'Neighborhood_Crawfor':
0.023708638284120738, 'Neighborhood_Edwards': -0.035128022877237 75, 'Neighborhood_Gilbert': -
0.04322378540096668, 'Neighborhood_IDOTRR': 0.0020914743868862755, 'Neighborhood_MeadowV': -
0.012436957138701042, 'Neighborhood_Mitchel': -0.01851829877752612, 'Neighborhood_NAmes': -
0.03417160414158187, 'Neighborhood_NPkVill': -0.015667208033143695, 'Neighborhood_NWAmes':
0.0003196326979913422, 'Neighborhood_NoRidge': 0.041087992467921094, 'Neighborhood_NridgHt':
0.037877822294122794, 'Neighborhood_OldTown': -0.000999004522480749, 'Neighborhood_SWISU': -
0.011082458602202309, 'Neighborhood_Sawyer': -0.021242242434213885, 'Neighborhood_SawyerW': -
0.007454599862953996, 'Neighborhood_Somerst': 0.03336002457672151, 'Neighborhood_StoneBr':
0.018376624218830948, 'Neighborhood_Timber': 0.002936017113491276, 'Neighborhood_Veenker':
0.021476191308147374, 'Condition1_Artery': -0.0064618917640027335, 'Condition1_Feedr': -
0.014316635164526026, 'Condition1_Norm': 0.031162491184031 44, 'Condition1_PosA':
0.0009035349756018403, 'Condition1_PosN': -0.006990955484372567, 'Condition1_RRAe': -
0.010200041771074517, 'Condition1_RRAn': 0.0032168653914965077, 'Condition1_RRNe': 0.0,
'Condition1_RRNn': 0.0026866326328465902, 'Condition2_Artery': 0.0, 'Condition2_Feedr':
0.0018026551497120967, 'Condition2_Norm': 0.026867950016556827, 'Condition2_PosA': 0.0,
'Condition2_PosN': -0.02867060516626876, 'Condition2_RRAe': 0.0, 'Condition2_RRAn': 0.0,
'Condition2_RRNn': 0.0, 'BldgType_1Fam': 0.023652999184162617, 'BldgType_2fmCon':
0.0005233725064872767, 'BldgType_Duplex': 0.005787566149885183, 'BldgType_Twnhs': -
0.022390601466083647, 'BldgType_TwnhsE': -0.00757333637445207, 'HouseStyle_1.5Fin': -
0.000464896691224299, 'HouseStyle_1.5Unf': 0.005608323028445502, 'HouseStyle_1Story':
0.009380716663524812, 'HouseStyle_2.5Fin': 0.007204028105112028, 'HouseStyle_2.5Unf': -
0.008584438071670997, 'HouseStyle_2Story': -0.004814635011024634, 'HouseStyle_SFoyer': -
```

0.005472914797291202, 'HouseStyle_SLvl': -0.002856183225871069, 'RoofStyle_Flat':
0.005158958593505576, 'RoofStyle_Gable': -0.012716751061363513, 'RoofStyle_Gambrel':
0.002503172898694793, 'RoofStyle_Hip': 0.005142094576952445, 'RoofStyle_Mansard': -
0.004496905617524708, 'RoofStyle_Shed': 0.004409430609734735, 'RoofMatl_ClyTile': -
0.018677817728019182, 'RoofMatl_CompShg': -0.0005792089721555786, 'RoofMatl_Membran': 0.0,
'RoofMatl_Metal': 0.0, 'RoofMatl_Roll': 0.0, 'RoofMatl_Tar&Grv': 0.003782121289759917,
'RoofMatl_WdShake': 0.014664199420433639, 'RoofMatl_WdShngl': 0.0008107059899810909,
'Exterior1st_AsbShng': 0.001432535743568182, 'Exterior1st_AsphShn': 0.0,
'Exterior1st_BrkComm': 0.0, 'Exterior1st_BrkFace': 0.05634928706238129, 'Exterior1st_CBlock':
0.0, 'Exterior1st_CemntBd': 0.010386536519811928, 'Exterior1st_HdBoard': -
0.025426271070352387, 'Exterior1st_ImStucc': 0.0, 'Exterior1st_MetalSd':
0.0004433441616177196825, 'Exterior1st_Plywood': -0.015019906974990402, 'Exterior1st_Stone':
0.010270345926154147, 'Exterior1st_Stucco': -0.009362072492041325, 'Exterior1st_VinylSd': -
0.002662474056784946, 'Exterior1st_Wd Sdng': -0.01745715336399257, 'Exterior1st_WdShing': -
0.008954171455526424, 'Exterior2nd_AsbShng': -0.00382001272162023, 'Exterior2nd_AsphShn': 0.0,
'Exterior2nd_Brk Cmn': -0.0019063932673700477, 'Exterior2nd_BrkFace': 0.017375744922547225,
'Exterior2nd_CBlock': 0.0, 'Exterior2nd_CmentBd': 0.010386536519811923, 'Exterior2nd_HdBoard':
-0.008071439029641447, 'Exterior2nd_ImStucc': 0.011204218942598779, 'Exterior2nd_MetalSd': -
0.004161917223200764, 'Exterior2nd_Other': 0.0, 'Exterior2nd_Plywood': -0.009494965233261845,
'Exterior2nd_Stone': -0.012029383782204167, 'Exterior2nd_Stucco': -0.016577904979073327,
'Exterior2nd_VinylSd': 0.0046646283948710504, 'Exterior2nd_Wd Sdng': 0.023680628709518627,
'Exterior2nd_Wd Shng': -0.011249741252975802, 'MasVnrType_BrkCmn': -0.010195397120561119,
'MasVnrType_BrkFace': 0.007981712323069243, 'MasVnrType_None': -0.004336459681190652,
'MasVnrType_Stone': 0.01391647260562233, 'MasVnrType_nan': -0.007366328126939774,
'Foundation_BrkTil': 0.012867319161544957, 'Foundation_CBlock': -0.0006434801680887047,
'Foundation_PConc': 0.014044631504371725, 'Foundation_Slab': -0.011367781452342263,
'Foundation_Stone': -0.005981148962943816, 'Foundation_Wood': -0.008919540082543,
'BsmtExposure_Av': -0.008650919473691168, 'BsmtExposure_Gd': 0.02985187173008206,
'BsmtExposure_Mn': 0.007032149503625778, 'BsmtExposure_No': -0.014355823043303385,
'BsmtExposure_nan': -0.01387727871671217, 'Heating_Floor': -0.0005681402017273736,
'Heating_GasA': -0.002611223343400934, 'Heating_GasW': 0.009419466645464567, 'Heating_Grav': -
0.0032217285582272215, 'Heating_OthW': 0.0, 'Heating_Wall': -0.003018374542109055,
'CentralAir_N': -0.012199625613152527, 'CentralAir_Y': 0.012199625613152503,
'Electrical_FuseA': -0.0021915004880258765, 'Electrical_FuseF': 0.00038490990887762817,
'Electrical_FuseP': 0.0019064608501257135, 'Electrical_Mix': 0.0, 'Electrical_SBrkr': -
9.9870270977177947e-05, 'Electrical_nan': 0.0, 'BsmtFullBath_0': -0.032249217465013386,
'BsmtFullBath_1': 0.014356839648310443, 'BsmtFullBath_2': 0.008667326256534113,
'BsmtFullBath_3': 0.00922505156016877, 'BsmtHalfBath_0': -0.003545702542576446,
'BsmtHalfBath_1': 0.0035457025425763918, 'BsmtHalfBath_2': 0.0, 'FullBath_0':
0.004631967745650905, 'FullBath_1': -0.04279531527236475, 'FullBath_2': 0.012972926966447034,
'FullBath_3': 0.02519042056026751, 'HalfBath_0': -0.012326130614986771, 'HalfBath_1':
0.024007811741380015, 'HalfBath_2': -0.0116816811263935 46, 'BedroomAbvGr_0':
0.004631967745650903, 'BedroomAbvGr_1': -0.012826371947321103, 'BedroomAbvGr_2': -
0.0026898683662901954, 'BedroomAbvGr_3': -0.006236595692253416, 'BedroomAbvGr_4':
0.026701560109025104, 'BedroomAbvGr_5': -0.004613660968359703, 'BedroomAbvGr_6': -
0.004967030880450732, 'BedroomAbvGr_8': 0.0, 'KitchenAbvGr_0': 0.0, 'KitchenAbvGr_1':
0.017220146310255206, 'KitchenAbvGr_2': -0.01379916130535596, 'KitchenAbvGr_3': -
0.003420985004899141, 'TotRmsAbvGrd_10': 0.049898562920572626, 'TotRmsAbvGrd_11': -
0.00844063439172549, 'TotRmsAbvGrd_12': -0.018677817728019235, 'TotRmsAbvGrd_14': 0.0,
'TotRmsAbvGrd_3': -0.027540586568179044, 'TotRmsAbvGrd_4': -0.029002266362666323,
'TotRmsAbvGrd_5': -0.016296102549476055, 'TotRmsAbvGrd_6': -0.00606052557664927,
'TotRmsAbvGrd_7': 0.016328204789896886, 'TotRmsAbvGrd_8': 0.0006057379061855914,
'TotRmsAbvGrd_9': 0.039185427560060646, 'Functional_Maj1': 0.003228418113989556,
'Functional_Maj2': 0.0021565479573871955, 'Functional_Min1': -0.008781127249399839,
'Functional_Min2': -0.015112414946035341, 'Functional_Mod': 0.0009374742085125562,
'Functional_Sev': 0.0, 'Functional_Typ': 0.017571101915545825, 'Fireplaces_0': -
0.022179955121966372, 'Fireplaces_1': -0.005248460513095744, 'Fireplaces_2':
0.03791584749543519, 'Fireplaces_3': -0.010487431860372643, 'GarageType_2Types': -
0.005240474021790891, 'GarageType_Attchd': 0.006927868293676182, 'GarageType_Basment':
0.0026858965186787234, 'GarageType_BuiltIn': 0.025032112760508794, 'GarageType_CarPort': -
0.003339206599097442, 'GarageType_Detchd': -0.011751752377315527, 'GarageType_nan': -
0.014314444574658533, 'GarageFinish_Fin': 0.015885118213897854, 'GarageFinish_RFn':
0.006982664604830484, 'GarageFinish_Unf': -0.008553338244070452, 'GarageFinish_nan': -
0.014314444574658535, 'GarageCars_0': -0.014314444574658533, 'GarageCars_1': -
0.039933741683028656, 'GarageCars_2': -0.012999957681962323, 'GarageCars_3':
0.06724814393964937, 'GarageCars_4': 0.0, 'MiscFeature_Gar2': 0.001970016424635145,
'MiscFeature_Othr': 0.0, 'MiscFeature_Shed': -0.007623150306262518, 'MiscFeature_TenC': 0.0,
'MiscFeature_nan': 0.005653133881627709, 'SaleType_COD': 0.0005658018332905322,
'SaleType_CWD': 0.022875556409550562, 'SaleType_Con': 0.011156924859350045, 'SaleType_ConLD':
-0.002647057594910522, 'SaleType_ConLI': 0.0, 'SaleType_ConLw': 0.0, 'SaleType_New': -
0.004357114934440299, 'SaleType_Oth': 0.0, 'SaleType_WD': -0.027594110572840263,
'SaleCondition_Abnorml': -0.003941675189968779, 'SaleCondition_AdjLand': -
0.0009651253335096546, 'SaleCondition_Alloca': 0.014385225510095748, 'SaleCondition_Family': -
0.007989835540838881, 'SaleCondition_Normal': 0.002868525488662256, 'SaleCondition_Partial': -
0.004357114934440298}

When you double up Lasso Alpha value to 0.002, the important parameters that emerged out are:

{'OverallQual': 0.29132185709681, 'YearRemodAdd': 0.05035348471336665, 'BsmtQual': 0.025224326114631945, 'BsmtFinType1': 0.026520437880475735, 'FireplaceQu': 0.07730351946562322, 'WoodDeckSF': 0.00592415857012477, 'LotShape_IR1': 0.00265153860058425, 'LandSlope_Mod': 0.046979156162293, 'Neighborhood_ClearCr': 0.010685056304988226, 'Neighborhood_NoRidge': 0.004073641124224836, 'Neighborhood_NridgHt': 0.0035426841090509275, 'Neighborhood_Somerst': 0.0017655567719418446, 'Condition1_Norm': 0.031117697587620362, 'BldgType_1Fam': 0.04946269718115428, 'Exterior1st_BrkFace': 0.07320932291662839, 'Foundation_PConc': 0.014749223562600 85, 'BsmtExposure_Gd': 0.020700958908036575, 'HalfBath_1': 0.03702163649413156, 'BedroomAbvGr_4': 0.03149875845411175, 'TotRmsAbvGrd_10': 0.046770059531114835, 'TotRmsAbvGrd_7': 0.002898267792853341, 'TotRmsAbvGrd_9': 0.018765879788899698, 'Fireplaces_2': 0.030931221651389285, 'GarageType_Attchd': 0.016147185525796515, 'GarageType_BuiltIn': 0.003104729875742302, 'GarageFinish_Fin': 0.006033618337523935, 'GarageFinish_RFn': 0.0024575520859224263, 'GarageCars_2': 0.000707358765869806, 'GarageCars_3': 0.12543125952038078}

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: I will go with Lasso Regularization as it makes the equation much simpler without making any compromise with other quality parameters.

The Ridge and Lasso Regularization outcome for my analysis:

Ridge Regression:

R-Square for Training Set: 0.9096859889725758

R-Square for Test Set: 0.8189583410011175

RSS for Training Set: 1.490729385003399

RSS for Test Set: 7.577743745473288

MSE for Training Set: 0.05979038895607673

MSE for Test Set: 0.08811402672916242

Lasso Regression:

R-Square for Training Set: 0.8723206736948281

R-Square for Test Set: 0.8156853791734743

RSS for Training Set: 2.107483893310431

RSS for Test Set: 7.714737993955898

MSE for Training Set: 0.0710909150973423

MSE for Test Set: 0.08890694385670009

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: After observing the high impact 5 attributes are: **OverallQual, GarageCars_3, Exterior1st_BrkFace, TotRmsAbvGrd_10, Neighborhood_NoRidge**

I removed the above ones from incoming data set and rebuild the LASSO model again. This time the following are 5 important predictor variables:

LandSlope_Mod, FireplaceQu, BsmtQual, GarageArea, GrLivArea

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?

Steps to make sure the model is robust and more generalizable:

1. The model should be simpler and should not be very high in variance. As this would mean overfitting and hence chance of performing bad with unseen data.
2. The following step needs to be taken to capture the variance:
    a. Outlier treatment – remove outliers to have dependable data set
    b. Appropriate Encoding of predictor variables - Ordinal and Categorical should be done.
    c. Impute missing values appropriately with Mean/Median/Mode and please do keep domain knowledge in mind when we do this operation.
3. Observe the following attributes and patterns to make sure the model is reliable and robust:
    a. Mean Square Error
    b. Residual Sum of Squares
    c. R2 Value
    d. Adjusted R2 Value
    e. F-Statistic value
    f. Residual plot to detect any patterns
    g. Make sure the error the normalized distribution.

Implications:

1. The above-mentioned metrics has significant impact on the model if not observed and necessary action/revision of model not taken:
    a. It may perform very well in test data, but poor with unseen data
    b. We need to proper trade-off between bias/variance and error residual as this combination is very essential and should align with business problem you are trying to solve. The metrics like (pertain to business problem we try to solve):
        i. Accuracy
        ii. Specificity
        iii. Sensitivity.
    c. Hence the model is built based on business case like Telecom Churn or Identifying Cancer patients may need different boundary values for the above metrics and error tolerance.