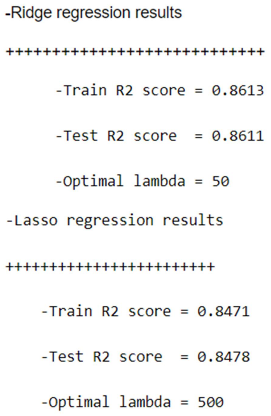
1) What is the optimal value of alpha for ridge and lasso regression? What will be the changes in the model if you choose double the value of alpha for both ridge and lasso? What will be the most important predictor variables after the change is implemented?
   Answer:
   Optimal alpha value of ridge = 50
   Optimal alpha value of lasso = 500

   Model co-efficients changes highly with doubling of the alpha value.
   This results in change of R2 score on train data from 0.8613 to 0.8485 and on test data from 0.8611 to 0.8559 for ridge regression. R2 score on the train data change from 0.8471 to 0.8238 and on test data from 0.8478 to 0.8345

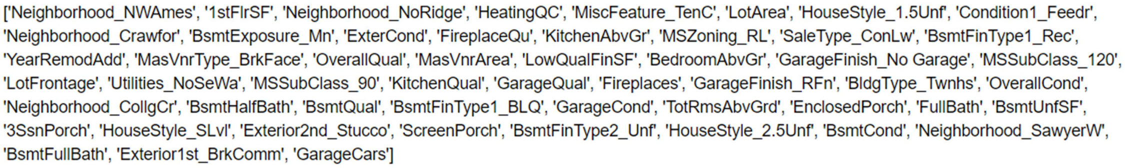
   Most important predictor remains same as "1stFlrSF" but its coefficient changes from 16388.5 to 12409 for Ridge regression
   Most important predictor remains change from "Neighborhood_NWAmes" to "HeatingQC" with coefficient of 30631.65 and 17801.3 respectivelu for Lasso regression

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:
   Ridge regression to be used for the prediction as the R2 score on train and test data is better than lasso



```
-Ridge regression results

++++++++++++++++++++++++++++

      -Train R2 score = 0.8613

      -Test R2 score  = 0.8611

      -Optimal lambda = 50

-Lasso regression results

++++++++++++++++++++++

    -Train R2 score = 0.8471

    -Test R2 score  = 0.8478

    -Optimal lambda = 500
```

   Lasso regression to be used for the interpretation of the influence of the predictors on the dependent variable as this method eliminates the feature by forcing the coefficient to 0.
   Below are the relevant coefficient from lasso in the order of the prominence



```
['Neighborhood_NWAmes', '1stFlrSF', 'Neighborhood_NoRidge', 'HeatingQC', 'MiscFeature_TenC', 'LotArea', 'HouseStyle_1.5Unf', 'Condition1_Feedr',
'Neighborhood_Crawfor', 'BsmtExposure_Mn', 'ExterCond', 'FireplaceQu', 'KitchenAbvGr', 'MSZoning_RL', 'SaleType_ConLw', 'BsmtFinType1_Rec',
'YearRemodAdd', 'MasVnrType_BrkFace', 'OverallQual', 'MasVnrArea', 'LowQualFinSF', 'BedroomAbvGr', 'GarageFinish_No Garage', 'MSSubClass_120',
'LotFrontage', 'Utilities_NoSeWa', 'MSSubClass_90', 'KitchenQual', 'GarageQual', 'Fireplaces', 'GarageFinish_RFn', 'BldgType_Twnhs', 'OverallCond',
'Neighborhood_CollgCr', 'BsmtHalfBath', 'BsmtQual', 'BsmtFinType1_BLQ', 'GarageCond', 'TotRmsAbvGrd', 'EnclosedPorch', 'FullBath', 'BsmtUnfSF',
'3SsnPorch', 'HouseStyle_SLvl', 'Exterior2nd_Stucco', 'ScreenPorch', 'BsmtFinType2_Unf', 'HouseStyle_2.5Unf', 'BsmtCond', 'Neighborhood_SawyerW',
'BsmtFullBath', 'Exterior1st_BrkComm', 'GarageCars']
```

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:

If the variables after creating the dummy variables being created are missing then following variables are first 5 important variables

Neighborhood_NPkVill, MiscFeature_Shed, Neighborhood_NoRidge, HouseStyle_1.5Unf, LotFrontage

If the entire variables are missing then 5 most important variables are as follows

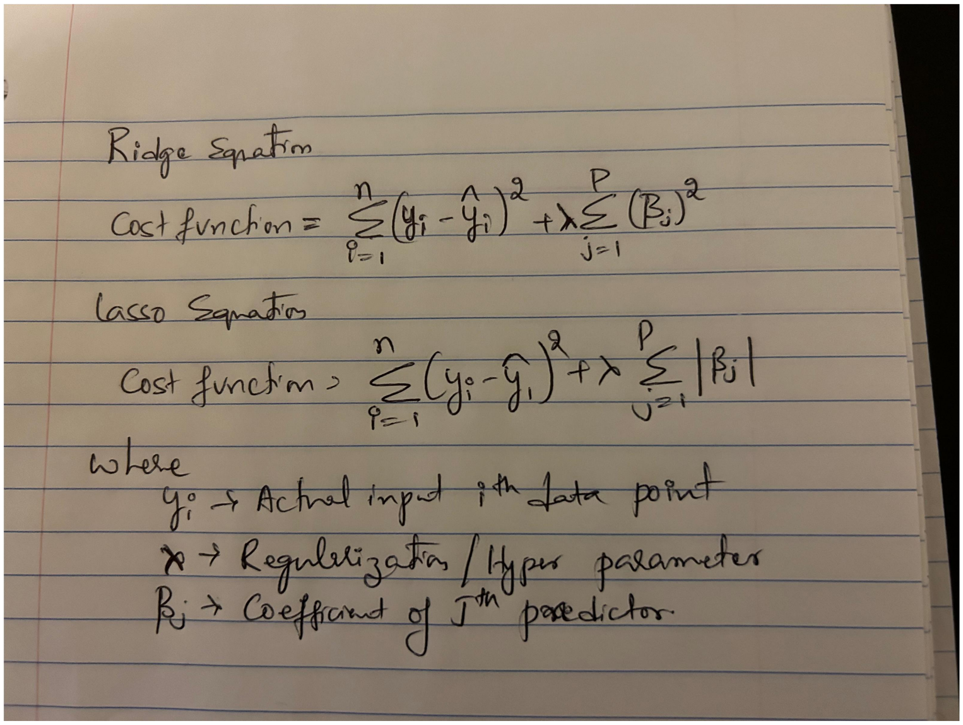
Fence_No Fence, HouseStyle_1.5Unf, LotFrontage, BsmtUnfSF, BsmtCond

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:

Model can be made robust and generalizable by achieving a trade-off between bias and variance. In other words, by preventing over fit resulting from complex model and under fit resulting for simplest model. Complexity of the model can be explained differently for different hypothesis class. For linear class this can be the degree of the linear model, value of the coefficients, decimal precision of the coefficients etc. This complexity is controlled by a concept called regularization wherein the models are penalized for it to be complex yet understand the patterns of the data.

Ridge and Lasso regularization methods can be used to penalise the model for its complexity by having an additional term in the cost function as given below to minimize and therefore calculate the resulting coefficient which is as simple as acceptable



Ridge Equation

Cost function $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} (\beta_j)^2$

Lasso Equation

Cost function $= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$

where

y_i → Actual input ith data point

λ → Regularization / Hyper parameter

β_j → Coefficient of Jth predictor

Regularization/hyper parameter (lambda) controls the complexity where high lambda value means that model is under fit and low lambda value means model is overfit. Optimal value is found using iterative method to find a value that gives low bias on the training set

Implication of the regularization on accuracy is that a portion of accuracy is compromised on the training data to achieve low variance