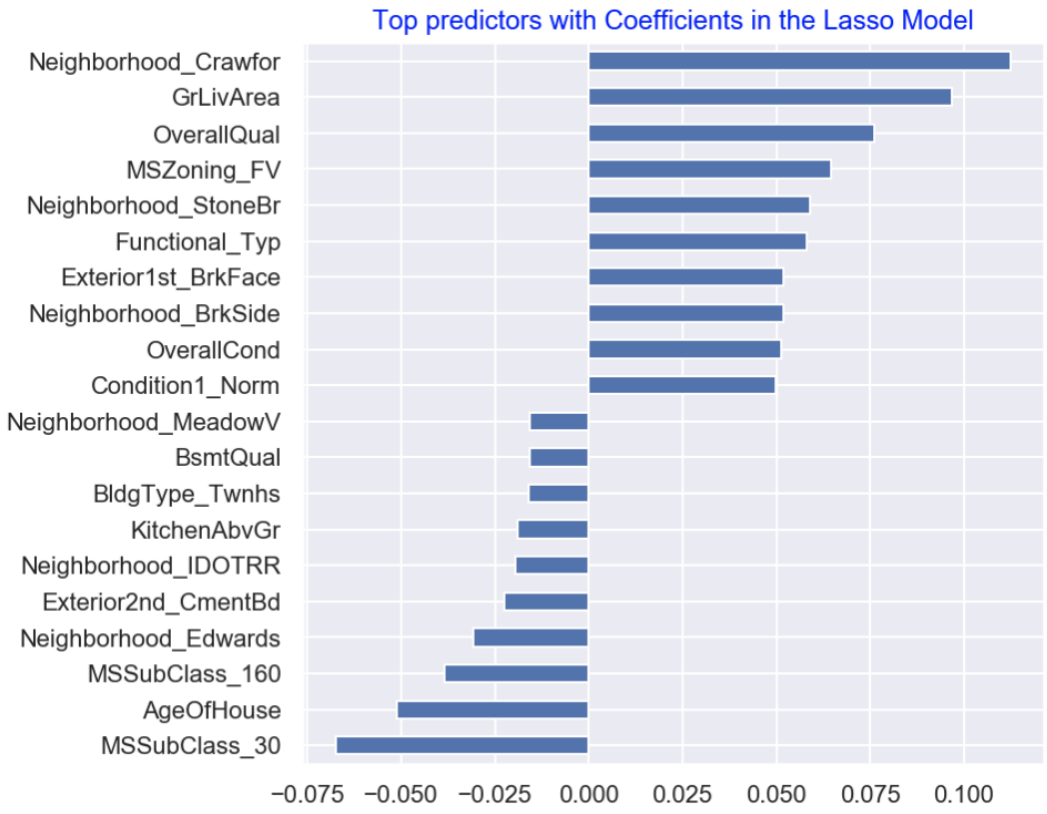
Question 1

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?

Answer 1

The optimal value of alpha for Ridge is 19.6973 and that of Lasso is 0.001



- The top 5 predictors are:
  - Neighborhood_Crawfor (Positive)
  - GrLivArea (Positive)
  - OverallQual (Positive)
  - MSSubClass_30 (Negative)
  - MSZoning_FV (Positive)
- Neighborhood_Crawfor is the top positive predictor and MSSubClass_30 is the top negative predictor

The metrics of final model are:

```
1 result_df
```

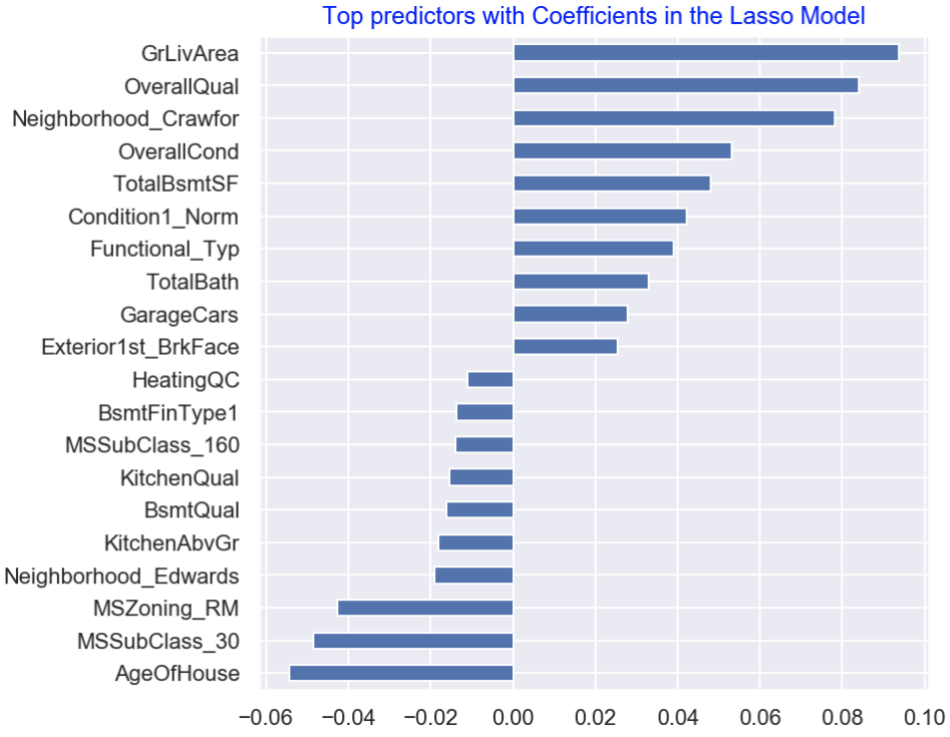
	model	alpha	r2_score_train	r2_score_test	rmse_train	rmse_test
0	LinearRegression	NaN	0.93865	-1.4683e+22	3.08588e+10	9.94977e+10
1	LassoRegression	0.001	0.91387	0.914597	0.114839	0.122165
2	RidgeRegression	19.6973	0.921593	0.91108	0.109569	0.124655
3	ElasticNetRegression	0.00126827	0.918873	0.914194	0.122453	0.122453

```
1 result_df.loc[1]
```

```
model              LassoRegression
alpha                        0.001
r2_score_train             0.91387
r2_score_test             0.914597
rmse_train                0.114839
rmse_test                 0.122165
Name: 1, dtype: object
```

When alpha value is doubled in Lasso Regression model

- r2 score of train and test data reduced by 0.011 and 0.122 respectively
- RMSE of train and test data increased by 0.007 and 0.004 respectively
- r2score of train data is 0.903 and r2score of test data is 0.909
- RMSE of train data is 0.122 and or test data is 0.126

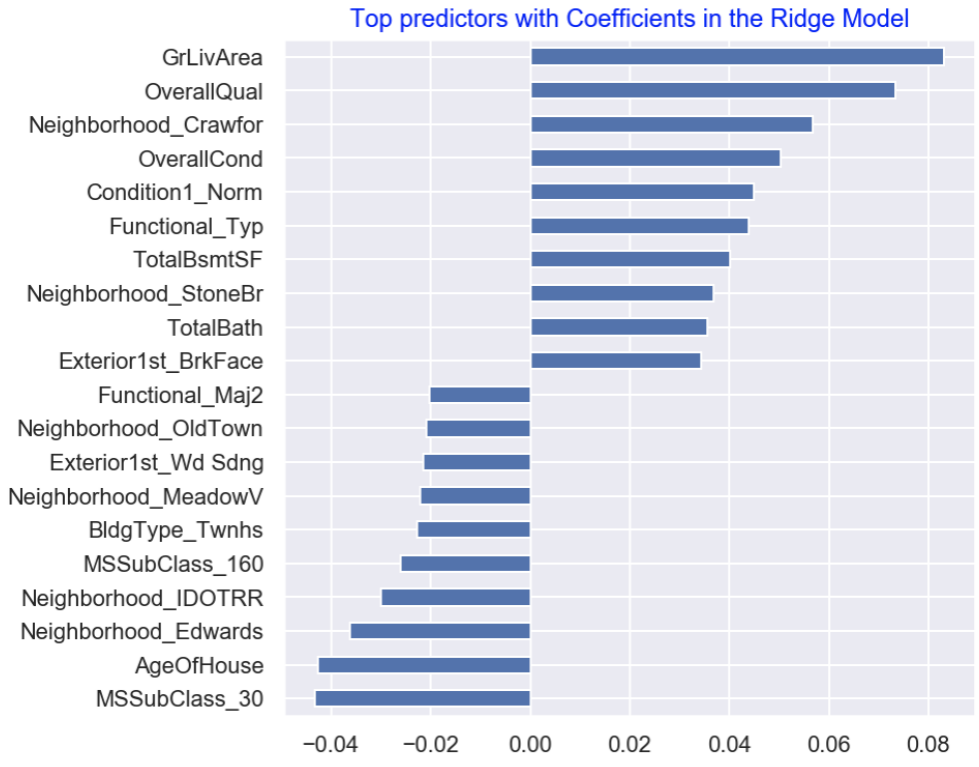


GrLivArea is the top positive predictor and AgeOfHouse is the top negative predictor when alpha value is doubled in Lasso Regression model

- The top 5 predictors are:
  - GrLivArea (Positive)
  - OverallQual (Positive)
  - Neighborhood_Crawfor (Positive)
  - AgeOfHouse (Negative)
  - OverallCond (Positive)

When alpha value is doubled in Ridge Regression model

- r2 score of train and test data has reduced by 0.006 and 0.0003 respectively
- RMSE of train and test data has increased by 0.0038 and 0.0002 respectively
- r2score of train data is 0.916 and r2score of test data is 0.911
- RMSE of train data is 0.113 and or test data is 0.125



- GrLivArea is still the top positive predictor and MSSubClass_30 is the top negative predictor when alpha value is doubled in Ridge Regression model
- However now, OverallQual is the second top positive predictor followed by Neighborhood_Crawfor and AgeOfHouse if the top second negative predictor followed by Neighborhood_Edwards
- The top 5 predictors are:
  - GrLivArea (Positive)
  - OverallQual (Positive)
  - Neighborhood_Crawfor (Positive)

- OverallCond (Positive)
- Condition1_Norm (Positive)

Conclusion and Recommendations when alpha value is doubled for Lasso and Ridge Regression
- Ridge preforms better than Lasso model when alpha value is doubled
- test data r2 score of Ridge model is 0.911 while that of Lasso is 0.909
- test data rmse of Ridge model is 0.125 while that of Lasso is 0.126
- With this we conclude that when alpha value is doubled, Ridge performs better and therefore we recommend Ridge model as a better fit in this scenario
- The top 5 predictors are:
  - GrLivArea
  - OverallQual
  - Neighborhood_Crawfor
  - OverallCond
  - Condition1_Norm

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 2

```
1 result_df
```

	model	alpha	r2_score_train	r2_score_test	rmse_train	rmse_test
0	LinearRegression	NaN	0.93865	-1.4683e+22	3.08588e+10	9.94977e+10
1	LassoRegression	0.001	0.91387	0.914597	0.114839	0.122165
2	RidgeRegression	19.6973	0.921593	0.91108	0.109569	0.124655
3	ElasticNetRegression	0.00126827	0.918873	0.914194	0.122453	0.122453

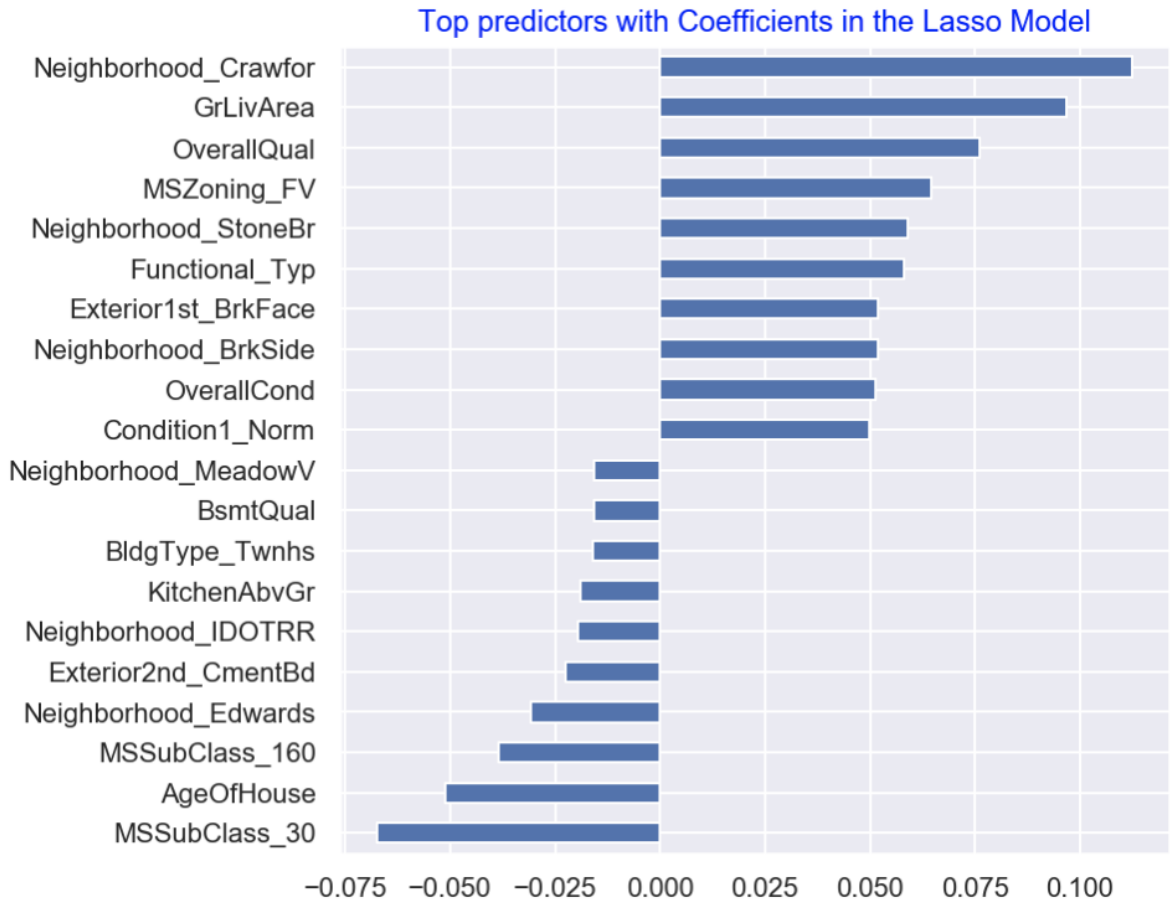
- Lasso Regression model performs better on train and test data with r2 score almost similar on train and test data depicting good fit on data
- Also RMSE on test data is the lowest for Lasso
- Based on these observations, we choose Lasso model and will recommend to apply Lasso Regression model to predict House Sales Price

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 3

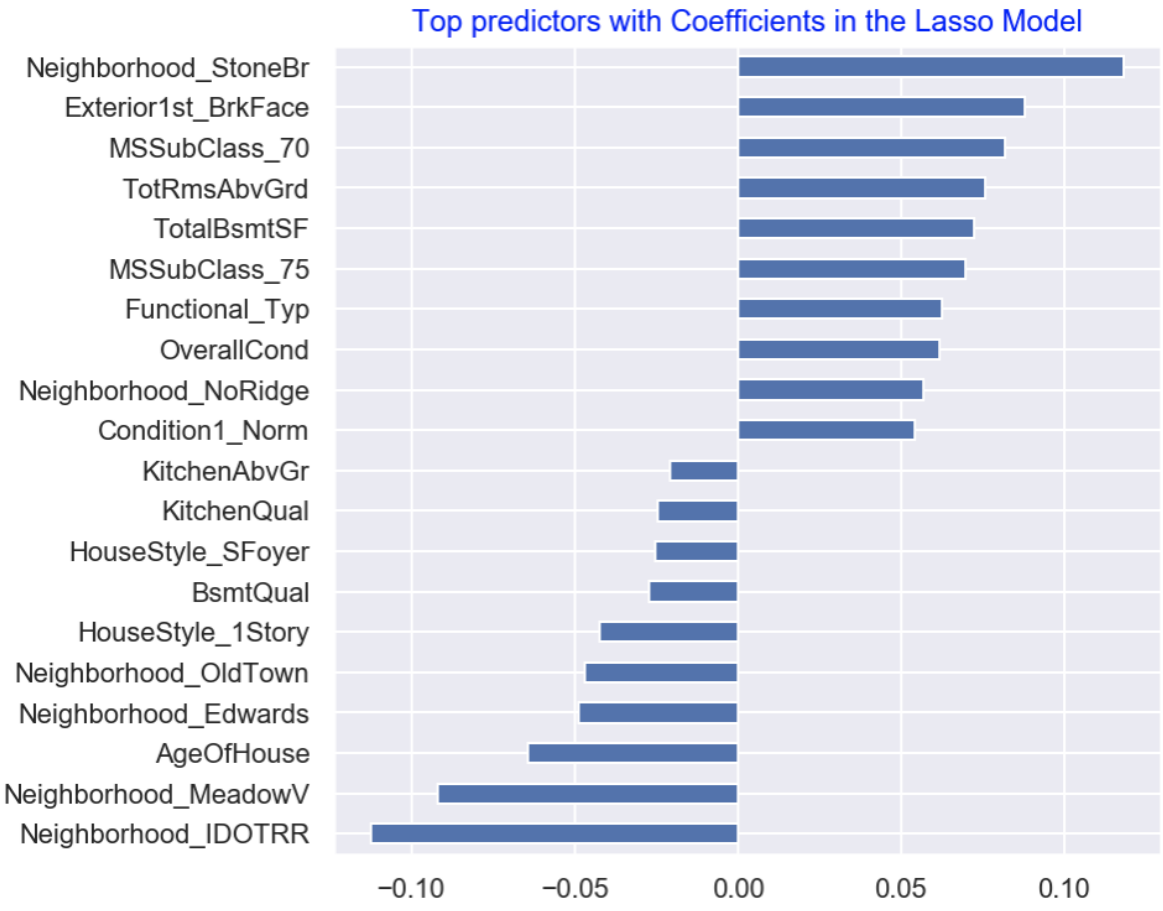
From final model – Lasso Regression



- Top 5 predictors are:
  - Neighborhood_Crawfor (positive)
  - GrLivArea (positive)
  - OverallQual (positive)
  - MSSubClass_30 (negative)
  - MSZoning_FV (positive)

- Since these top predictors are not available in incoming data so we need to drop these columns from the data set and try to build the model again and check for the top predictors

- When top 5 predictors are dropped then in Lasso Regression model
  - r2 score of train and test data has reduced by 0.019 and 0.032 respectively
  - RMSE of train and test data has increased by 0.012 and 0.143 respectively
  - r2score of train data is 0.895 and of test data is 0.882
  - RMSE of train data is 0.127 and of test data is 0.143



- Top 5 predictors are

- Neighborhood_StoneBr (positive predictor)
- Neighborhood_IDOTRR (negative predictor)
- Neighborhood_MeadowV (negative predictor)
- Exterior1st_BrkFace (positive predictor)
- MSSubClass_70 (positive predictor)

Question 4

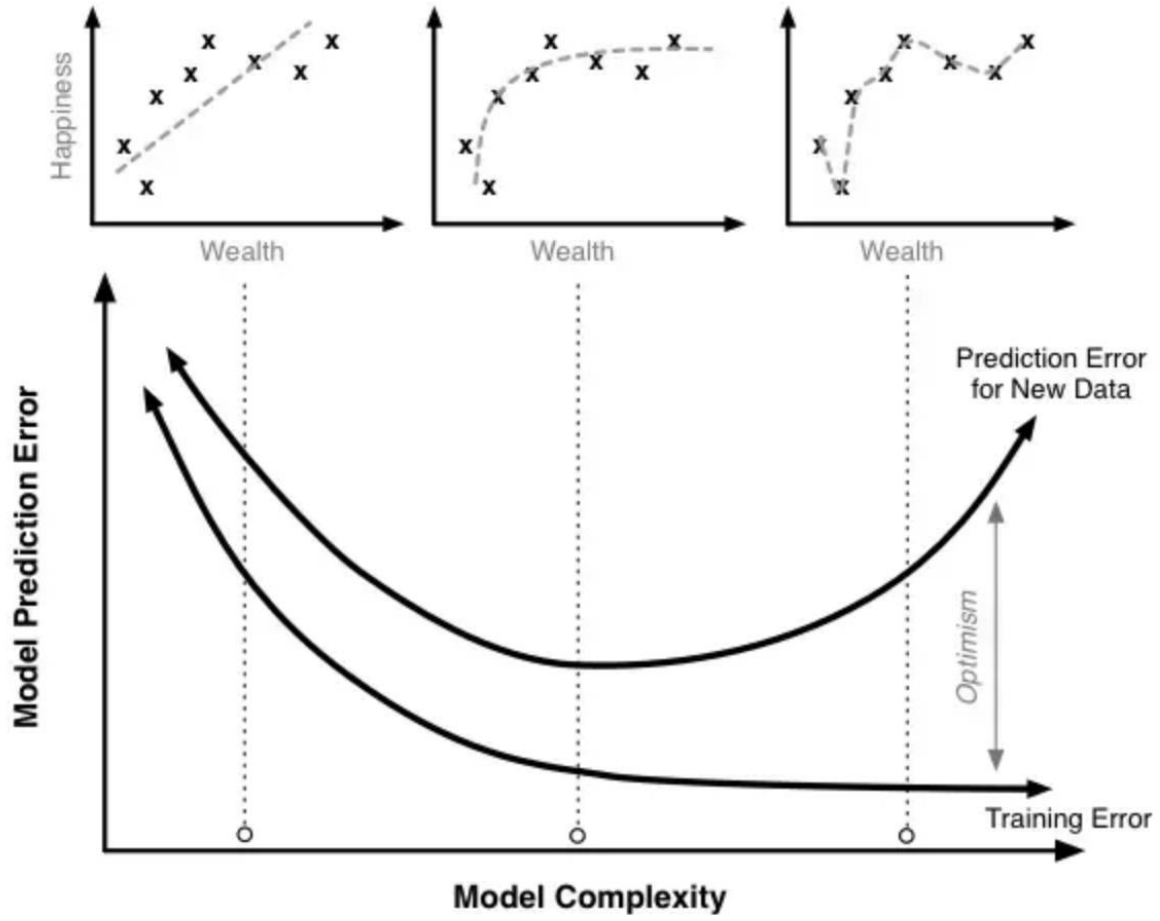
How can you make sure that a model is robust and generalisable? What are the implications of the same for the accuracy of the model and why?

Answer 4

Generalization of model in machine learning is defined as the model's ability to react to new data. A machine learning algorithm is used to fit a model to data. For this purpose, model is presented with Training data having multiple observations. The model learns from the training data. Once the model learns from the training data, the model is applied on unseen data for predictions.

Now here's where problems can emerge. If model is over trained on the training data, then it will memorise all the data points or observations of training data resulting is almost zero training error and very high accuracy, but will fail miserably when applied on the new data.

We then say that the model is incapable of generalizing, or that it is overfitting the training data.
This problem of overfitting and underfitting can also be explained with bias-variance trade-off
Bias is known as error in the learning algorithm. It occurs when algorithm is unable to learn much from the training data (underfitting). Problem of variance occurs when algorithm over-learns from training data and tries to fit each and every point of training data, there by missing the pattern in data and fails on test data (overfitting)
Bias-Variance trade off is a technique by which model is trained in such a way that it captures the regularities in data enough to get reasonably accurate model which at the same time is generalizable to a different set of data points (test data) from same data source. This is called having optimum bias and optimum variance



In above example, from left to right, models have been trained longer on trained data. The training error curve in the bottom shows that the training error reduces when model complexity increases and when model is trained more on training data. The training error is the most (high bias, low variance) for the simplest model where the model hits very less actual data points, it's the lowest for the highest complex model (low bias, high variance) where model hits every data point. While in the middle, the model hits most of the data points(not exactly all) and the Training error is moderate (optimal bias and optimal variance). Also when these models are applied on test data, the test error is highest for the simplest model. It's again highest for the most complex model while its lowest for the middle complex model

With this we can conclude that, to create good predictive models in machine learning the model complexity should be such that the model should not overfit the data. The accuracy of the model should be such that it should fit both train and test data in a way resulting in minimum train and test error at the same time

Bias Variance Trade off can be achieved by following methods

1. Regularization – It's the method to reduce overfitting by penalising the coefficients of features, to get high accuracy for both training and test data (e.g. Lasso, Ridge, ElasticNet)
2. By building more complex model – to resolve underfitting problem by building complex model and to resolve overfitting problem by bringing in more data with addition of regularization
3. K-Fold Cross-Validation method – Here data is divided into k-sets. The first set or first fold is the validation data set, and the first fold is removed from the total number of folds(where, suppose k=10). For each iteration, one fold is taken for validation and the remaining is used for training. In every iteration, the test fold is different. This method is effective yet requires huge computational power