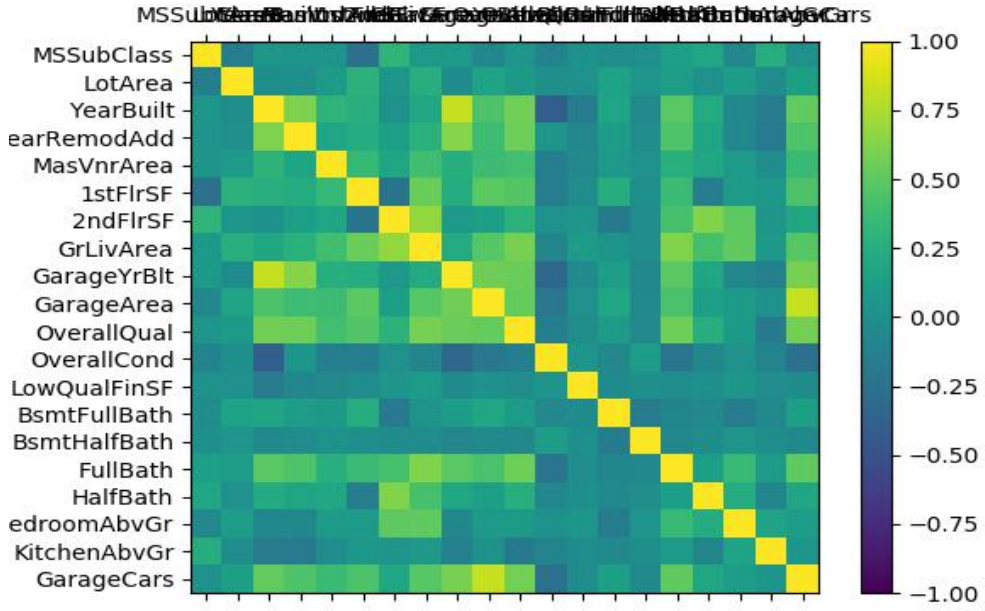
1) For Data pre-processing I have done three things

a) removing NAN values

b) converting 'object' values into numeric values by using Label Encoder

C) normalizing Data

d) selecting features using below graph it is based on covariance values.




Problems faced:- For converting 'object' data into numeric ,first I used the method of making dummies columns similar to one hot encoding. It worked well but when I had done same thing on test.csv it generated different no. Of columns because of different type of data in same column and therefore could not fit into my classifier .But later with Label Encoding it worked well.

2) I tried with two types of classifier

a) Random Forest

b) SVM

For this first we evaluated best parameters and then for real test we have trained on full tarin.csv but for choosing best out of two we divided data into two parts with 35% to test data and 65% to train data and evaluated the accuracy as follows:

RandomForest = 91.7%

Svm=89.2%

Therefore choose Random Forest for real test data.

3) The hyper-parameters of Random forest and Svm were tuned using Grid Search and K-Fold with cv=5.

a)Random Forest:-

Tuning Parameters:-

‘bootstrap': [True,False],
'max_depth': [5,6],
'max_features': ['sqrt','auto'],
'min_samples_leaf': [3, 4, 5],
'min_samples_split': [7,9],
'n_estimators': [60,70,80]

From these best parameters were used and a classifier was trained for getting accuracy of test data.

Best parameters:'max_features': 'sqrt', 'n_estimators': 80, 'min_samples_leaf': 4, 'max_depth': 6, 'bootstrap': False, 'min_samples_split': 7

B)SVM:-

Tuning Parameters:-

'C': [100,400,800],
'kernel': ['linear','poly','rbf'],
'degree': [3,4,5],
'gamma': [0.1,0.01,0.001]

From these best parameters were used and a classifier was trained for getting accuracy of test data.

Best parameters:'C': 800, 'gamma': 0.001, 'degree': 3, 'kernel': 'rbf'