CS 418: Introduction to Data Science

Exploratory Data Analysis

Fall 2018

Reena Mary Puthota

rputho2@uic.edu

Sai Sharan Nagulapalli

snagul2@uic.edu

**Problem Statement:**

Given the following datasets:

1. Ask a home buyer to describe their dream house and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. This dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence
2. It also predicts the category of the price - high, medium high, medium low, low

**Data Cleaning:**

**For the numerical values:**

We handled missing data for LotFrontage, and MasVnrArea by filling them with the modes of the respective columns. We did so, because unlike other columns that follow, NA wasn’t already a predefined label for the data.

For example, GarageType has a predefined value of NA that means that it isn’t applicable, as the house doesn’t have a garage.

**For Categorical Values:**

For some of the columns ['GarageType', 'GarageYrBlt', 'GarageFinish', 'GarageCars', 'GarageArea', 'GarageQual','GarageCond'] that already have a category that specifies NA - None or No Garage etc we found that all those columns had missing values corresponding to their group. Implying that there is no Garage and hence no values for Garage Condition, Garage Quality etc.

For these columns we have filled all the NaN values with their respective NA values

**Cross Validation :** We are using a KFold cross validation with **5 Folds**

**Data Science Solutions :**

**Feature Selection:**

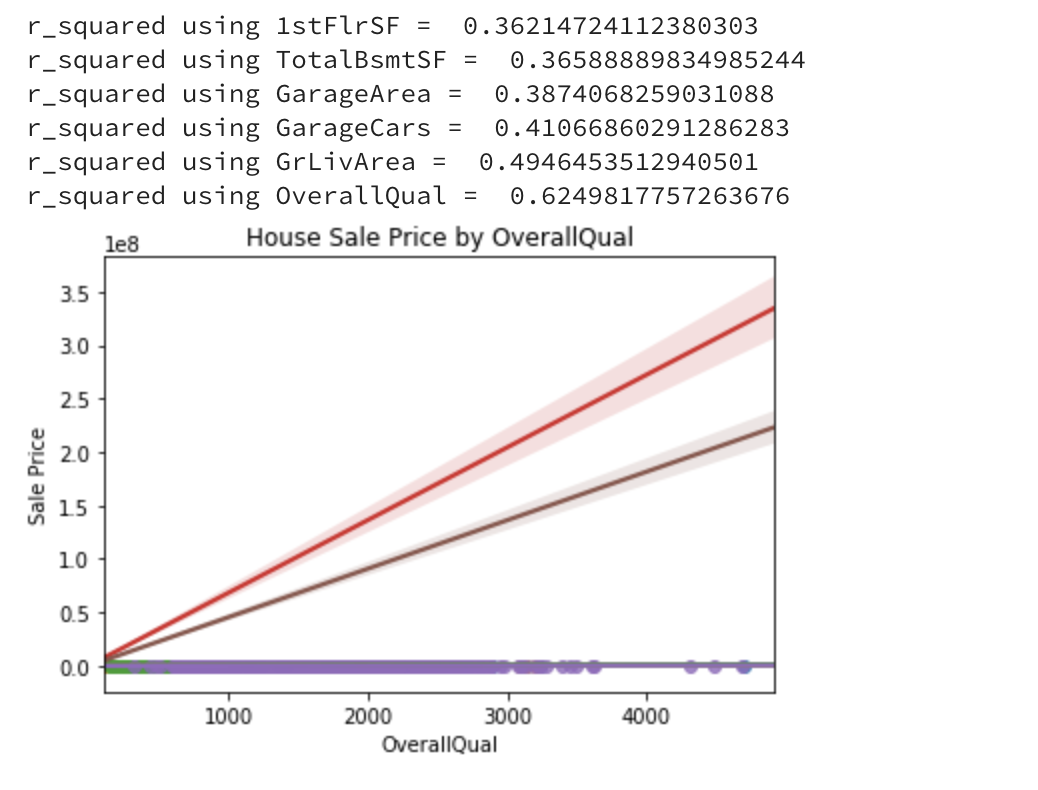
We have computed the correlation coefficients of all the features against the dependent variable i.e SalePrice and selected the top 6 columns as our selected feature set.

**Regression Techniques:**

**Single Linear Regression:**

We have performed Linear Regression with each of the selected predictor variables against the dependent variable and plotted the results on a line chart as follows:

It shows that mostly only the top two parameters are influencing the Sale Price more predominantly than many others



**Multi Linear Regression:**

We have performed Multi Linear Regression with various combinations of the selected predictor variables list against the dependent variable and the top 3 results on based on r-square and adjusted\_r\_square values are as follows:

Showing that the OverQuality of the House is dominant factor of all the predictor variables and has the most influence



**Classification Techniques:**

**Decision Tree:**

We are using the GridSearchCV for trying classifiers for different combinations of the parameters of the Classifier and also multiple combinations of the Feature

**GridSearchCV(DecisionTreeClassifier(), param\_grid = parameter\_grid, cv = kfold)**

**parameter\_grid =**

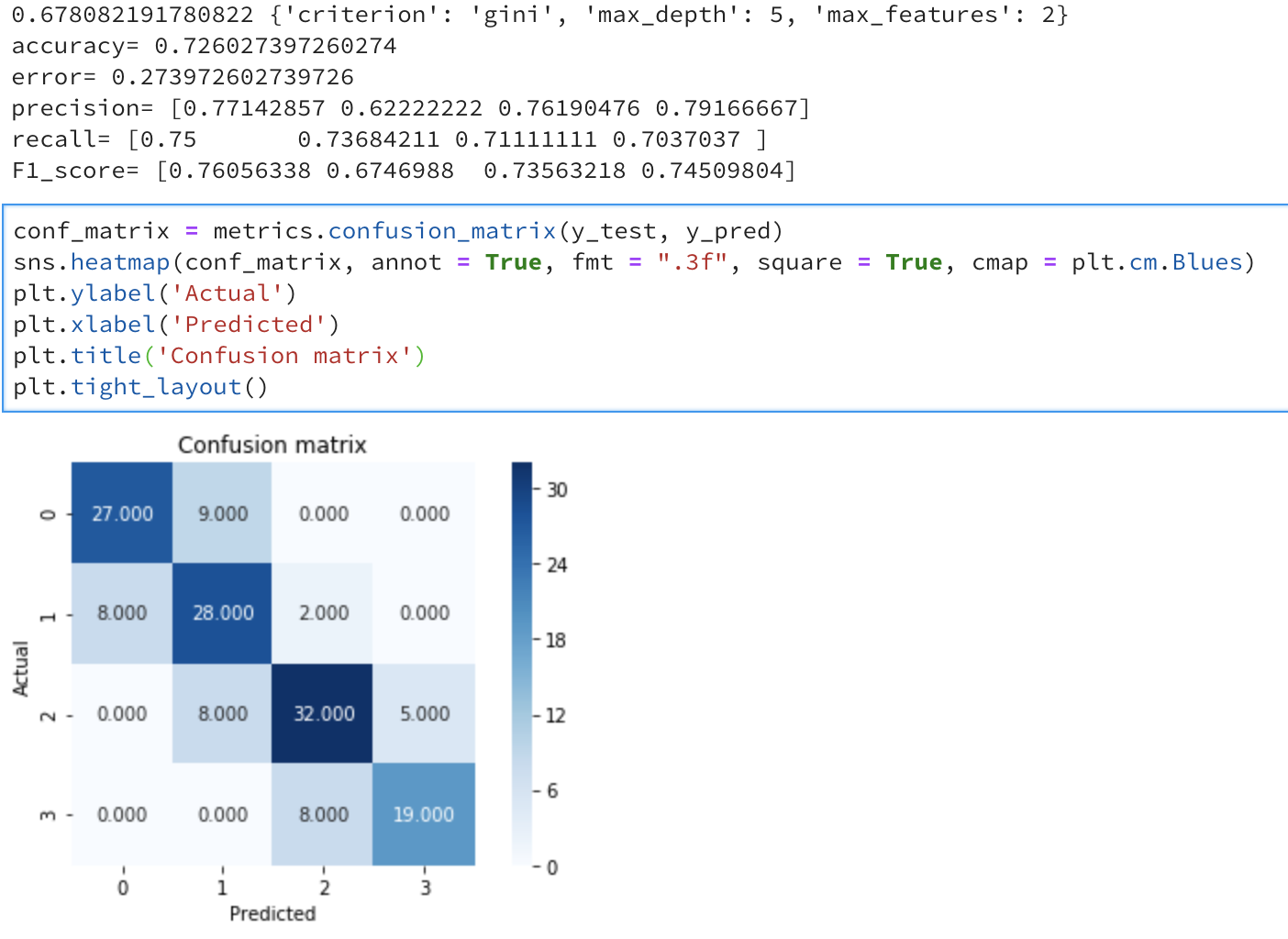
**{'criterion': ['entropy','gini'],'max\_depth': [1, 2, 3, 4, 5], 'max\_features': [1, 2, 3, 4]}**

We found the following results with best accuracy being :

**0.6815068493150684**

With the model parameters being **{'criterion': 'gini', 'max\_depth': 5, 'max\_features': 3}**

**Then the model is tested against a test split from the train data and Results are represented as a confusion matrix as follows below**

****

**KNeighborsClassifier - No feature selection**

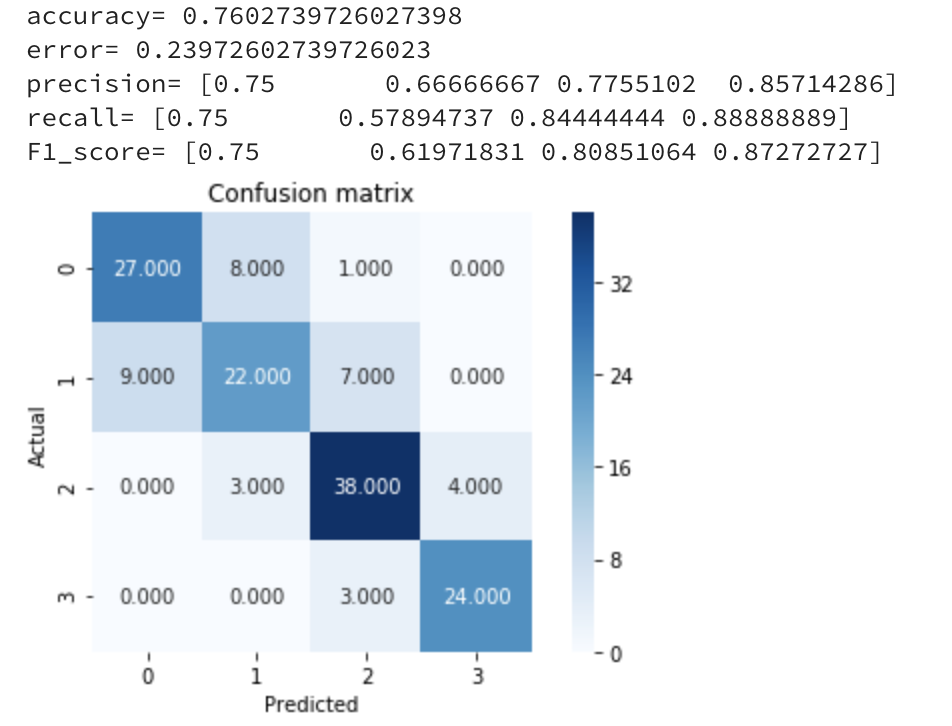
We have tried a different way here , implemented an approach to iterate through the splits of the KFold, perform modelling on each split and compare the results against the current top classifier based on accuracy.

Once the iteration has finished we display the top classifiers from each Fold Iteration and save the best among them.

Here are the results from the above analysis for the top classifier:

TOP

Neighbour: 7  
Accuracy 0.7054794520547946



**KNeighborsClassifier - With feature selection of top 6 columns with highest correlation coefficient**

We follow the same iterative approach but this time we have altered the feature to the top six columns based on correlation coefficient and it seems to improve the results slightly. But the test run on a split of data from the train was producing less efficient results than the classifier without any feature selection

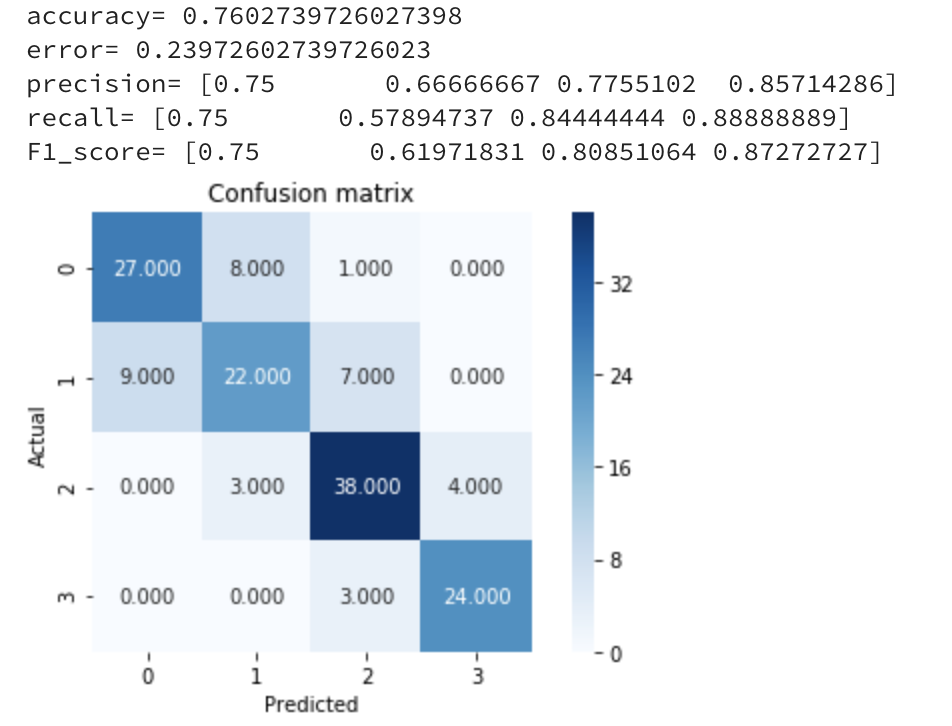
The Results are as follows :

TOP :

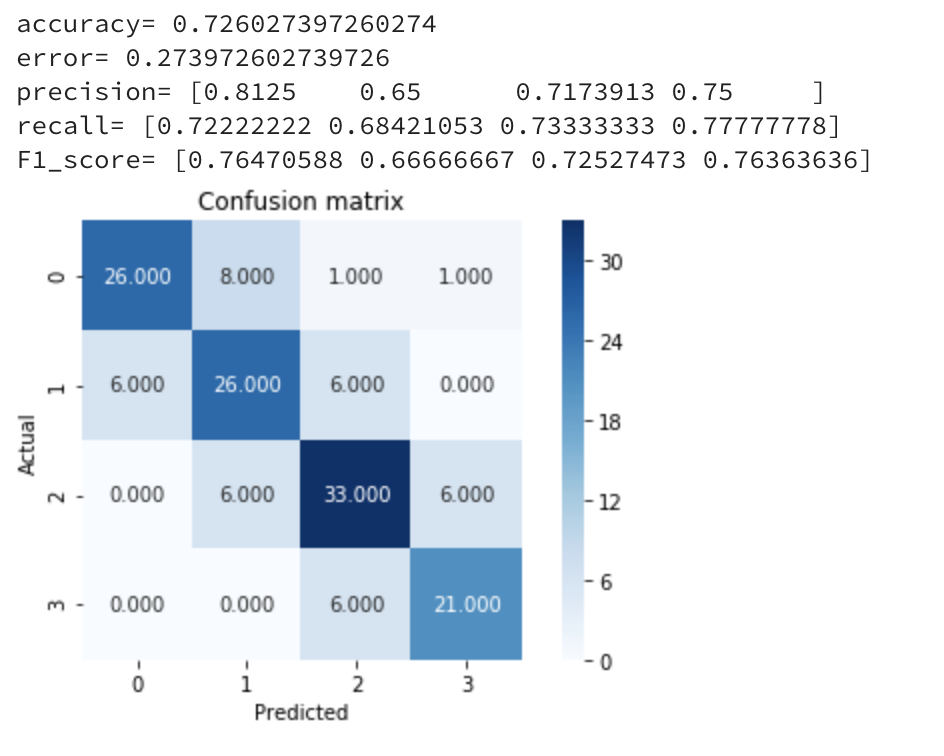
**Best K-fold result:**

Neighbour: 6  
Accuracy 0.7636986301369864

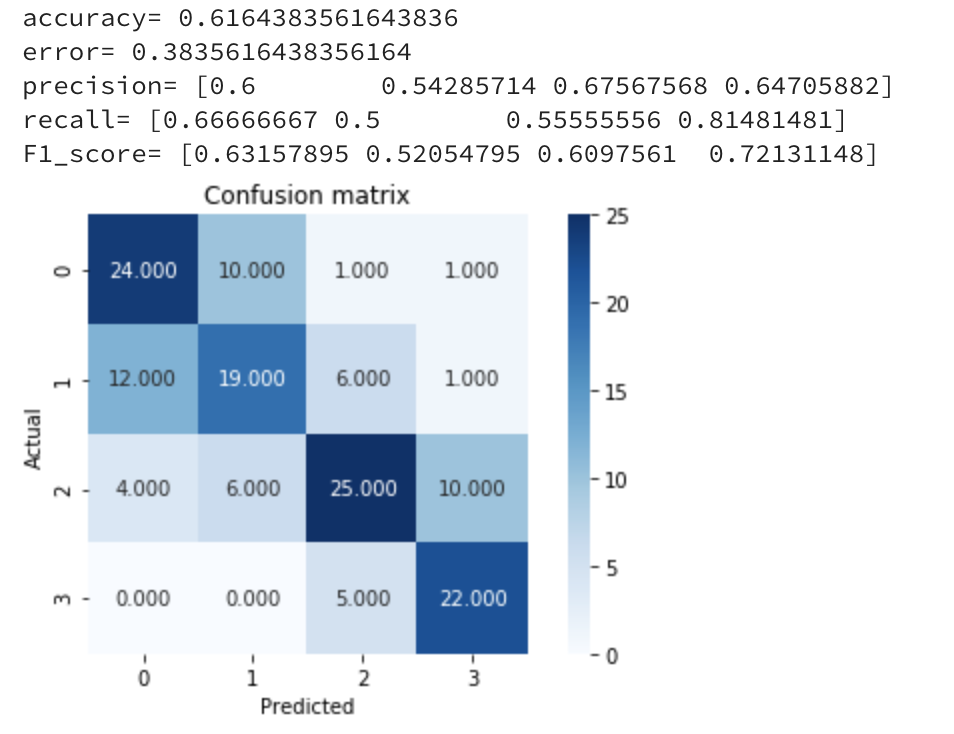
TEST: **(K=6)**



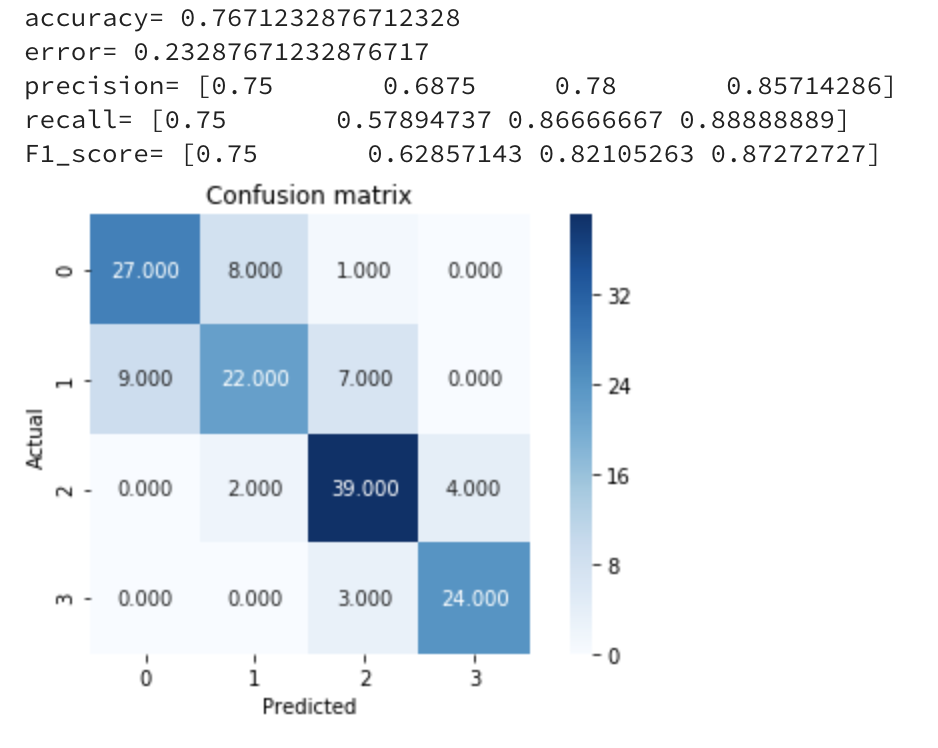
**GuassianNB - With feature selection of top 6 columns with highest correlation coefficient**

****

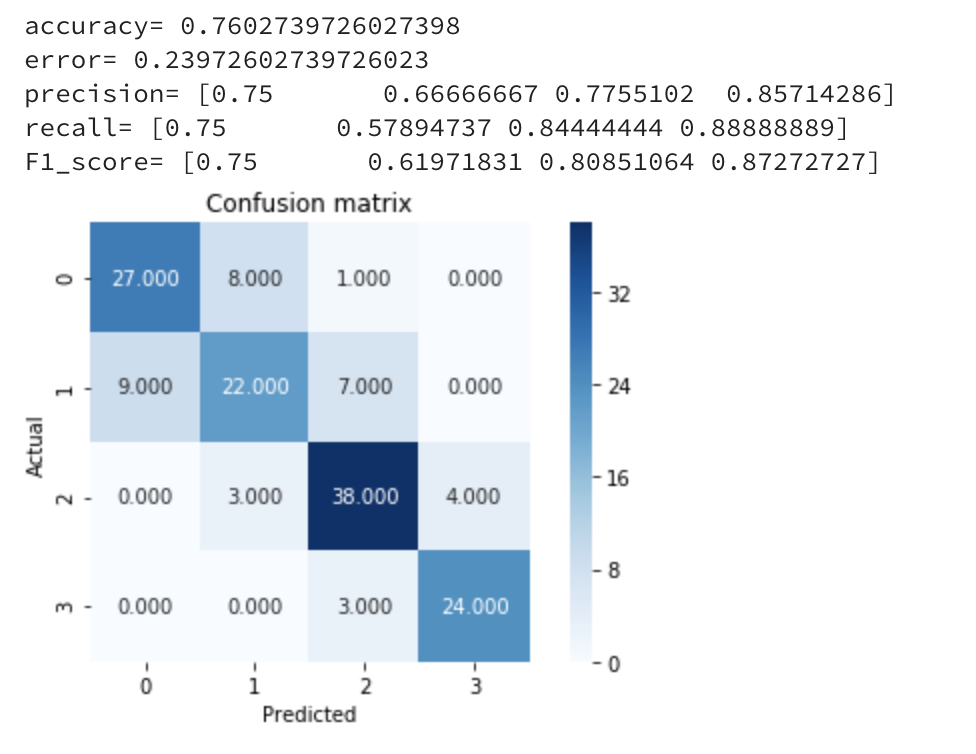
**MultinomialNB - With feature selection of top 6 columns with highest correlation coefficient**

****

**Linear SVM - With feature selection of top 6 columns with highest correlation coefficient**

****

**RBF Kernel SVM**

****

**Conclusion**

The top factors affecting the price of a House are Overall Quality, Total Basement SF, Living Area, Garage Area etc

With the primary influencer being Overall Quality

But Some interesting findings are that the 1st Floor Area influences more than the Basement Area