A logo with black text

Description automatically generated

CS770 Machine Learning

Assignment1: Linear Regression, Ridge& lasso Regression

09/25/2024

Submitted by,

Shyam

Introduction:

In this assignment, we have explored The Boston Housing Data set which was collected by the U.S census service to predict the median house price in Boston. The main aim of this assignment is to calculate the performance of different models using the regressor metrics such as R2, RMSE. The Machine Learning models we applied are Simple Linear Regression, Multi Linear Regression, Ridge, Lasso, and OLS Regression models.

Methods

This dataset has 506 rows and 14 features to predict/ calculate the Median House Price Value(MEDV). Below Regressor models are implemented:

1. Simple Linear Regression: A Feature which has highest correlation with our target variable MEDV.
2. Multi Linear Regression: Considering all the Feature variables with our target variable MEDV.
3. Ridge Regression: To avoid Overfitting, we perform Regularization. Ridge adds a apenalty as model complexity increases.
4. Lasso Regression: Creating a model to make sure it fits the data properly and reduce the loss by adding a penalty term(alpha).
5. OLS Regression model: Using OLS we can identify important features of the dataset.

Steps:

1. After loading the data set, checked for null values, datatype format, duplicated rows.
2. There are 6 columns not more than 4% of the null values, hence dropped those.
3. Performed EDA and got some interesting insights from the dataset.
   1. Plotted histograms, scatter plots, heat map on the dataset.

* There is a right skewness for several features like CRIM, ZN, LSTAT, DIS, and MEDV
* As they are right skewed, the values are nearer to 0, which tells us that:
  + CRIM is right skewed means, most of the houses in the dataset have low crime rates.
  + NOX is also right skewed means most of the houses have with high concentration of Nitric oxide.
  + DIS is also right skewed means most of the most areas are closer to the employment careers.
  + PTRATIO is also right skewed means most of the areas have moderate people teacher ratio.
  + LSTAT is right skewed means most of the houses have a lower percentage of lower status individuals.
* With the help of this scatter plots against MEDV:
  + Negative relation between crime rate and house prices, higher crime rate- lower house price.
  + Slightly negative relation between indus and medv, higher industrial activity- lower house price.
  + houses near to the charles river are slightly higher compared to the houses far from river.
  + Many points are around Age of 100, which means many houses are older than 100 years.
  + positive relation between dis and medv, housed near to the employment center are cheaper.

1. Checked for outliers using boxplots and using pandas clip function cap them with higher and lower fence values.
2. Applied transformations for several features where skewness is visible.
3. Handling Multicollinearity using VIF and dropping the feature with more VIF. After calculating VIF values we observed that tax has more multicollinearity than RAD, hence dropping TAX as well from features.
4. Split the dataset int training and testing.
5. Train and save the model as pkl files using joblib.
6. Load the pickle files and evaluate the metrics with different models.

Results and Discussion:

The performance of each Regressor model was evaluated using Mean Square Error, and r2 score.

Mean Squared Error: Measures the avg squared difference between the actual and predicted values.

R Squared: explains how close the data is fitted to regression line.

Below are the metrics:

Simple Linear Regression - MSE:  77.9891, R-squared: 0.075371

Linear Regression - MSE:  33.6327, R-squared: 0.601255

Conclusions:

In this assignment, we observed that Multi Linear regression has the best performance, both with R2 (0.60) and MSE(33.6). This indicates that Multiple linear regression model explains 60% of the variance in the target variable than simple linear model. It also says that all the features are necessary to increase the model performance.

Importance Of Feature Scaling:

* Feature Scaling is important in Regression analysis than the Calssification models because the target variable is numerical, and the features are also numerical.
* In this dataset some feature values are in 3 digits and some features are in 1 digit.
* If we do not do Feature Scaling, the feature with highest values will dominate the feature with lowest feature.
* We can achieve Feature scaling using scikit-learn library. Depending on the type of model we are training/ testing we will select either Standard Scaler or MinMax scaler.
* Standard scaler will shrinks the data from 0 to 1, while min max scaler allows us to tweak the min and max value of each feature.
* Feature scaling also helps to normalize the data such that there won't be any skewness. skewness will affect the model performance.
* In this assignment we have used MinMax Scaler because we are using Linear Regression model.