4 Modeling

After cleaning the data and analysing the various features, we can proceed with model building. Here we will use a supervised machine learning model and since the target variable is continuous we will use linear regression, decision tree and random forest for model building.

4.1 Linear Regression

For linear regression models we need to work with features that are strongly related with features , they may be proportionally or inversely related. We will use the value of correlational matrix for setting the limit on features which we want to use for model building and calculate the errors.

	Feature Name	Total Feature	Corr value	r2 score	MAE	MSE
0	['RM' 'PTRAT' 'LSTAT' 'MEDV']	3	0.5	0.488164	4.40443	41.678
1	['RM' 'LSTAT' 'MEDV']	2	0.6	0.488164	4.14244	37.3831
2	['LSTAT' 'MEDV']	1	0.7	0.488164	4.86401	46.3363
3	['INDUS' 'NOX' 'RM' 'TAX' 'PTRAT' 'LSTAT' 'MEDV']	6	0.4	0.488164	4.3945	42.6519

As we can see from the table above we are selecting values of correlational matrix which are 0.5, 0.6, 0.7,0.4. Basically selecting features which are strongly related and values which are not so strongly related and comparing their errors. R2 score remains constant throughout but for MAE and MSE we can see that they are optimum around the corr value of 0.6. If we select features which are highly correlated and features which are highly non correlated the errors increase. So it would be wise to use corr value of 0.6 to build a linear model

4.2 Decision Tree

We can also build a decision tree model on this problem. For the decision tree model we made a random search for the hyperparameters of the decision tree.. After getting the parameters values from random search,we did a grid search and passed every

value to parameters of the decision tree. And got the best parameters which gives the lowest error. Then we compared the test results with the default values of parameters of the decision tree with the best parameters and saw that the default parameter performed well. There may be various reasons for this one could be we did not give a broad range in random search as it would be too much computational load for the local machine.

4.3 Random Forest

In the random forest model we did the same thing as the above like the decision tree. We performed a random search then a grid search extracted the best parameters and compared the result with the default parameters. Here our random forest model performed better than the base model and we got an improvement of 2.18%.

4.4 Random Forest with feature improvement

In the above random forest model we used all the features.
But feature importance is a method where we can look at which features are important for model building in random forest.
From the figure we can see what weights each features. So after having a look at the importance value we decided to use the top 3 features to build a random forest.

We used the best value parameters from the above model and applied that to this model and we got a 0.58% from the base model.

	importance
RM	0.475831
LSTAT	0.337513
DIS	0.058791
CRIM	0.038557
NOX	0.018237
TAX	0.016091
PTRAT	0.015945
AGE	0.014654
B	0.012604
INDUS	0.006015
RAD	0.003753
ZN	0.001184
CHAS	0.000824