TASK 4

PROBLEM STATEMENT – Use the new Housing dataset provided to predict the house prices based on input data. You are required to try out various regression algorithms for this purpose.

COLUMN NAMES --

Variables in order:
CRIM per capita crime rate by town
ZN proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS proportion of non-retail business acres per town
CHAS Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX nitric oxides concentration (parts per 10 million)
RM average number of rooms per dwelling
AGE proportion of owner-occupied units built prior to 1940
DIS weighted distances to five Boston employment centres
RAD index of accessibility to radial highways
TAX full-value property-tax rate per $10,000
PTRATIO pupil-teacher ratio by town
B 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
LSTAT % lower status of the population
MEDV Median value of owner-occupied homes in $1000's

STEP 1 - Import the dataset and all required libraries.

STEP 2 - Preparing the dataset. Name the columns if required. Clean for any NaN values. Split the dataset into Train and Test sets.

STEP 3 – Decide on models you will use and create a array of it.

STEP 4 – Use KFold cross-val scoring mechanism to find the best model for yourself. You can use any scoring mechanism. ‘r2’ scoring is preferred.

STEP 5 - Finalize on the model with best score.

STEP 6 – Train the final model with training set and predict with the test set.

We know, this will be tough for most of you. For beginners we want them to understand the concepts and main document first before

trying this out. Those who have be through the basics can try a hand at this task. We will be explaining the solution of this task live tomorrow.