**BACKGROUND**

This Guided Project was initiated as part of Dataquest’s course entitled “Machine Learning Fundamentals”. The data used for the project was taken from Jeffrey C. Schlimmer’s Automobile Data Set. The original data set and its description can be downloaded from the [UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets/automobile). For this project, machine learning model used was K Nearest Neighbors (KNN) using scikit-learn’s built in commands for KNN predictive analysis.

**DESCRIPTION**

As with previous projects, this Guided Project started with loading the data into the Jupyter Notebook. Data Cleaning followed next with the following changes made on the data:

1. Data fields having “?” symbols instead of the correct data were replaced by NaN.
2. Columns (num\_doors and num\_cylinders) with numeric values spelled out were replaced by their equivalent floating point values.
3. All numeric values were converted to floating point values.
4. Rows with null values were dropped. Columns affected are num\_doors, peak\_rpm and horsepower.
5. Null values were replaced by the average of the that particular column. Columns affected are normalized\_losses, bore, stroke, and price.

After cleaning, the numerical data was normalized using MinMaxScaler().

A function was defined to train and test models for a single fold KNN model and multiple fold KNN model (knn\_train\_test). The single fold KNN model was extensively explored to find the optimal KNN model for the data set. The following cases were explored:

1. Single fold univariate models for the default value of n\_neighbors (n\_neighbors = 5).
2. Single fold univariate models for a variety of values of n\_neighbors (1, 3, 5, 7 and 9).
3. Single fold KNN models using:
   1. All feature columns;
   2. Best 2 features as determined from #1 using the default value of n\_neighbors;
   3. Best 3 features as determined from #1 using the default value of n\_neighbors;
   4. Best 4 features as determined from #1 using the default value of n\_neighbors;
   5. Best 5 features as determined from #1 using the default value of n\_neighbors;
   6. Best 2 features as determined from #1 using a variety of values of n\_neighbors;
   7. Best 3 features as determined from #1 using a variety of values of n\_neighbors;
   8. Best 4 features as determined from #1 using a variety of values of n\_neighbors; and
   9. Best 5 features as determined from #1 using a variety of values of n\_neighbors.

For the model optimization, the root mean square error was calculated to determine which models are performing better against the others.

Finally, the code for multiple fold KNN model was tested for several univariate models of default n\_neighbors value using 3 folds. An average error was calculated for each univariate model.

**FILES**

The following files are included in this project:

|  |  |
| --- | --- |
| FILE NAME | DESCRIPTION |
| GP\_Car\_Price\_Prediction.ipynb | The Jupyter Notebook version of the project. |
| Imports-85.data | The data set in \*.data format |
| GP\_Car\_Price\_Prediction.html | The project in html format for easy viewing |
| GP\_CARPRICE\_PREDICT\_README.docx | Short documentation for the project. |