**Creation of MLS Workspace:**

First, you will have to visit the Microsoft Machine Learning studio by copying and pasting the link, <https://studio.azureml.net>, into the search bar.

Then, you will click signup and choose the 8-hour trial, guest workspace option. This will lead you to this page:

A screenshot of a computer

Description automatically generated

Press the Experiments tab and select the new button at the bottom of the page, this will pop up a section containing types of samples, you press “Blank Experiment” which create a new workspace to perform machine learning.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

**Building Machine Learning Model Using Automobile Price data:**

Select the Automobile price dataset from the sample item, which is a subsection of the Saved Dataset Item, then drag the dataset to the workspace.

A screenshot of a phone

Description automatically generated

To perform EDA, you will click on the “Automobile price data (Raw) in the workspace, then choose “Visualize”. This will display all the variables in the data along with the summary statistics, if you click on each variable, it can either display a boxplot or a histogram for visualizations under the visualization tab. There are 206 rows and 26 columns. Here are some examples of Summary Statistics, Boxplots and Histograms:

**Summary Statistics:**

* Horsepower: Mean = 104.26, Median = 95, Minimum = 48, Maximum = 288, Standard Deviation = 39,71, Unique Values = 59, Missing Values = 2
* Wheel-Base: Mean = 98.76, Median = 97, Minimum = 86.6, Maximum = 120.9, Standard Deviation = 6.02, Unique Values = 53, Missing Values =
* Length: Mean = 174.05, Median = 173.1, Minimum = 141.1, Maximum = 208.1, Standard Deviation = 12.34, Unique Values = 74, Missing Values = 0
* Peak-rpm: Mean = 5125.37, Median = 5200, Minimum = 4150, Maximum = 6600, Standard Deviation = 479.33, Unique Values = 23, Missing Values = 2

There are 2 outliers in the Wheel-Based Boxplot, 4 outliers in the horsepower boxplot, 1 outlier for the length boxplot and there is 1 outlier for the peak-rpm boxplot.

* A screenshot of a graph

  Description automatically generatedA screenshot of a graph

  Description automatically generatedA screenshot of a graph

  Description automatically generatedA screenshot of a graph

  Description automatically generated

A screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generatedA screenshot of a graph

Description automatically generated

To clean missing data, In the experiment side bar, type in the search bar, “Clean Missing Data”, the item, called “Clean Missing Data” will pop up. Click and drag it to the workspace and connect it to the dataset, then press RUN which is at the bottom of the page. This will clean the Automobile Price data. We will be using the default values option for missing data; you can change that using the Clean Missing Data sidebar.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

To find the top variables with the highest correlation to our target variable, Price, you will use the Filter Based Feature Selection. To do that, search “Filter Based Feature Selection”, then click and drag the item to the workspace, and connect it to the Automobile price data. Then to select the target variable, click Launch Column Selector in the Filter Based Feature Selection sidebar on the right of the screen, then select the column, price, and click the grey check mark, then, click RUN. Lastly click Features, on the Filter Based Feature Selection item, to display the variables that have the highest correlation to price. We will be using the top 5 variables. The Pearson’s Correlation was used to determine the correlation of the variables.

A white background with black dots

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

To select columns of the Automobile price dataset for machine learning, search, “Select Columns in Dataset”, then click and drag the Select Columns in Dataset item to the workspace. Then, connect it to the Clean Missing Data item. Next, click Launch Column selector, use the top 5 variables that are highly correlated to price along with price, then press the grey button. Click RUN.

A blue and white rectangular object with a white background

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

To split the Automobile price data into a training and testing subset for machine learning, search, “Split Data”, then click and drag the Split Data item, connect it to the Select Columns in Dataset item. To create a 67% split, type “0.67” under “Fraction of rows in the firs…” and change the random seed to 7 in the Split Data sidebar to the right of the page. Click RUN.

A white background with black dots

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

To use the Linear Regression Model, search, “Linear Regression”, and select the Linear Regression item. Click and drag the item to the workspace and then make random seed to 7 using the Linear Regression sidebar on the right of the page. OLS was used to compute linear regression. Click RUN.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

To train the model, search, “Train Model” in the search bar, click and drag Train Model item to the workspace and connect it to the Linear Regression item and the Split Data item. Next, click on the launch column selector which is Train Model sidebar to the right of the page, and choose price as the target variable, click the grey check button. Click RUN.

A white background with black dots

Description automatically generatedA screenshot of a computer

Description automatically generated

A screenshot of a flowchart

Description automatically generated

To score the model, search, “Score Model”, then select the Score Model item, click and drag it to the workspace. Connect it to the Train Model item and Split Data item. Click RUN.

A screenshot of a computer

Description automatically generatedA screenshot of a diagram

Description automatically generated

To Evaluate the model, search, “Evaluate Model”, then select the Evaluate Model item, click and drag it to the workspace. Connect it to the Score Model item. Click RUN. To look at the results of the Evaluation Model, click on the item, and select visualize. The Coefficient of Determination is the r-squared.

A white background with black dots

Description automatically generatedA screenshot of a computer screen

Description automatically generated

A screenshot of a computer

Description automatically generated

To cross validate the model, search, “Cross Validate Model”, then select the Cross Validate Model, click and drag it to the workspace. Connect it to the Linear Regression item and the Select Columns in Dataset item. Select price as the target variable using Launch column selector in the Cross Validate Model sidebar, then click the grey check mark. Also change the Random seed to 7. Lastly, click RUN.

A screenshot of a phone

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated