Data Preparation

Data Preparation

Data preparation is a vital step in the machine learning pipeline. Just as visualization is necessary to understand the relationships in data, proper preparation or **data munging** is required to ensure machine learning models work optimally.

The process of data preparation is highly interactive and iterative. A typical process includes at least the following steps:

1. **Visualization** of the dataset to understand the relationships and identify possible problems with the data.
2. **Data cleaning and transformation** to address the problems identified. It many cases, step 1 is then repeated to verify that the cleaning and transformation had the desired effect.
3. **Construction and evaluation of a machine learning models**. Visualization of the results will often lead to understanding of further data preparation that is required; going back to step 1.

By the completion of this lab, you will:

1. Recode character strings to eliminate characters that will not be processed correctly.
2. Find and treat missing values.
3. Set correct data type of each column.
4. Transform categorical features to create categories with more cases and likely to be useful in predicting the label.
5. Apply transformations to numeric features and the label to improve the distribution properties.
6. Locate and treat duplicate cases.

Lab Steps

1. Make sure that you have completed the setup requirements as described in the Lab Overview section.
2. Now, run jupyter notebook and open the “DataPreparation.ipynb” notebook under Module 3 folder.

3. Examine the notebook and answer the questions along the way.

Question 1

1.0/1.0 point (graded)

What can you conclude about aggregating the hardtop and convertible categories to hardtop_convert?

- (●) It seems like a good idea because hardtop_convert category does appear to have values distinct from the other body style. ✔
- ( ) It seems like a bad idea because hardtop_convert category does appear to have values distinct from the other body style.
- ( ) It seems like a good idea because hardtop_convert category does NOT appear to have values distinct from the other body style.
- ( ) It seems like a bad idea because hardtop_convert category does NOT appear to have values distinct from the other body style.

Submit You have used 2 of 2 attempts

Question 2

1.0/1.0 point (graded)

From the scatter plots, it appears that the relationships between curb_weight and log_price and city_mpg and log_price are more linear, compared to the relationships between curb_weight and price and city_mpg and price respectively. What can you conclude from that?

- It is likely that curb_weight is better in predicting log_price than city_mpg.
- It is likely that curb_weight is better in predicting price than city_mpg.
- ◉ It is likely that curb_weight is better in predicting log_price than price. ✔
- It is likely that city_mpg is better in predicting log_price than curb_weight.

Submit You have used 2 of 2 attempts

Question 3

1.0/1.0 point (graded)

How many cases have duplicates?

- ◉ 12 ✔
- 22
- 1000
- 1012

Submit You have used 2 of 2 attempts