Analysis of a Car Sell Dataset Part 2

There are two goals in the second analysis task: (1), train linear regression models to predict the selling prices of cars; (2) assess the data ethics issues. There are *7* questions in this portfolio.

The first goal involves a standard Data Science workflow: exploring data, building models, making predictions, and evaluating results. In this task, we will explore the impacts of feature selections and different sizes of training/testing data on the model performance. We will use another cleaned car_sells sub-dataset that **is different from** the one in “Analysis of a Car Sell Dataset” task 1. This goal covers Questions *1-6*.

Question *7* is about data ethics issue.

```
your_name = "John xxx/Jane xxx"
your_student_id = "XXXXXXXX"

import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn import linear_model
from sklearn.metrics import r2_score

import seaborn as sns
import matplotlib.pylab as plt
%matplotlib inline
```

Q1: Import Cleaned Car Sell Dataset

The csv file named 'car_sells_clean_data.csv' is provided. You may need to use the Pandas method, i.e., `read_csv`, for reading it. After that, please print out its total length.

Q2: Explore the Dataset

- Use the methods, i.e., `head()` and `info()`, to have a rough picture about the data, e.g., how many columns, and the data types of each column.
- As our goal is to predict cars' selling prices given other columns, please get the correlations between year/km_driven/seller_type/fuel/owner and selling_price by using the `corr()` method.
- To get the correlations between different features, you may need to first convert the categorical features (i.e., seller_type and owner) into numerial values. For doing this, you may need to import `OrdinalEncoder` from `sklearn.preprocessing` (refer to the useful exmaples here)
- Please provide ***necessary explanations/analysis*** on the correlations, and figure out which are the ***most*** and ***least*** corrleated features regarding selling_price. Try to ***discuss*** how the correlation will affect the final prediction results, if we use these features to

train a regression model for selling_price prediction. In what follows, we will conduct experiments to verify your hypothesis.

Q3: Split Training and Testing Data

- Machine learning models are trained to help make predictions for the future. Normally, we need to randomly split the dataset into training and testing sets, where we use the training set to train the model, and then leverage the well-trained model to make predictions on the testing set.
- To further investigate whether the size of the training/testing data affects the model performance, please randomly split the data into training and testing sets with different sizes:
  - Case 1: training data containing 10% of the entire data;
  - Case 2: training data containing 90% of the entire data.
- Print the shape of training and testing sets in the two cases.

Q4: Train Linear Regression Models with Feature Selection under Cases 1 & 2

- When training a machine learning model for prediction, we may need to select the most important/correlated input features for more accurate results.
- To investigate whether feature selection affects the model performance, please select two most correlated features and two least correlated features regarding selling_price, respectively.
- Train four linear regression models by following the conditions:
  - (model-a) using the training/testing data in case 1 with two most correlated input features
  - (model-b) using the training/testing data in case 1 with two least correlated input features
  - (model-c) using the training/testing data in case 2 with two most correlated input features
  - (model-d) using the training/testing data in case 2 with two least correlated input features
- By doing this, we can verify the impacts of the size of traing/testing data on the model performance via comparing model-a and model-c (or model-b and model-d); meanwhile the impacts of feature selection can be validated via comparing model-a and model-b (or model-c and model-d).

Q5: Evaluate Models

- Evaluate the performance of the four models with two metrics, including MSE and Root MSE
- Print the results of the four models regarding the two metrics

Q6: Visualize, Compare and Analyze the Results

- Visulize the results, and perform ***insightful analysis*** on the obtained results. For better visualization, you may need to carefully set the scale for the y-axis.
- Normally, the model trained with most correlated features and more training data will get better results. Do you obtain the similar observations? If not, please ***explain the possible reasons***.

Q7: Data Science Ethics

*Please read the following examples Click here to read the example_1. Click here to read the example_2.

*Then view the picture My Image Please compose an analysis of 100-200 words that evaluates potential ethical concerns associated with the infographic, detailing the reasons behind these issues.