ECE 9063 Data Analytics Foundation

Assignment 1: Forecasting

**Student Name:** Jianping Ye

**Student Number:** 250887769

**Instructor:** Katarina Grolinger

**Problem Statement**

The used car market is a perfect place for finding cars in descent condition and with fair prices. It is also the reason that the market has ever-growing popularity. However, cars have diverse conditions and the market trend is not stationary all the time which makes the prices fluctuate. It will be beneficial for both buyers and sellers if we could make a model to predict the value of cars such that they can make a more confident decision of whether buying or selling. In this report, the forecasting problem is defined as follow: predict the price of a used car in the current year given a set of relevant attributes.

**Dataset Description**

Link to the data:

<https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes>

The folder mentioned above has many files for each of the car manufacturers. In this report, the dataset selected is “Audi.csv” which is the file specific for Audi cars. It contains 9 attributes and 10668 samples, and the attributes are listed below:

* Model: The model code of the car
* Year: registration year of the car
* Price: price on the market
* Transmission: type of gearbox, either manual, automatic, or semi-auto
* Mileage: distance used of the car
* fuelType: type of fuel the engine uses, either diesel, petrol, hybrid, or other
* tax: road tax
* mpg: miles per gallon
* engineSize: size of engine in litres

Noticeably, model, transmission, and fuel type have categorical data that needs to be transformed into numerical values. All the attributes in the dataset are considered in the model as they are all important factors contributing to the price of cars in the real-world business.

**Algorithms**

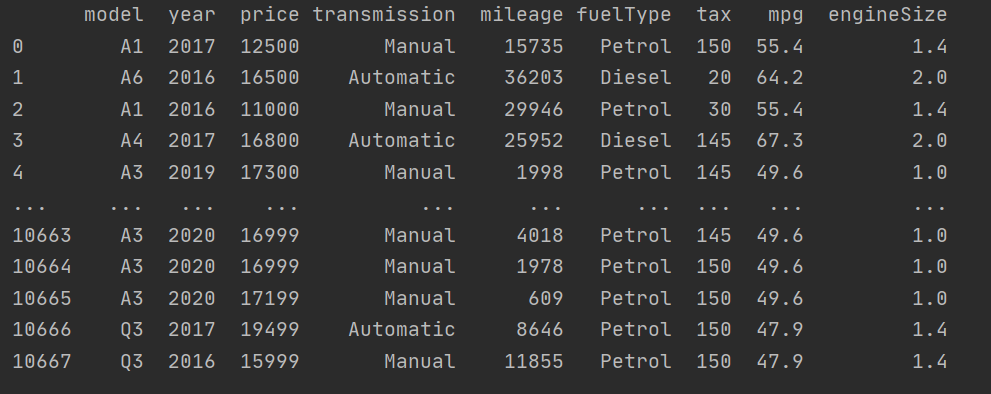
The first algorithm is multivariate linear regression. In this hypothesis, multiple independent variables contribute to the dependent variable. In the context of our dataset, multiple factors are affecting the price of cars and they all potentially have a linear relationship with the price, which make it a straight-forward and promising model.

The second algorithm is support vector regression. Support vector regression adheres to the basic principle of support vector machine, which is the maximum margin characteristic, but it is used for regression instead. The kernel selected is linear kernel.

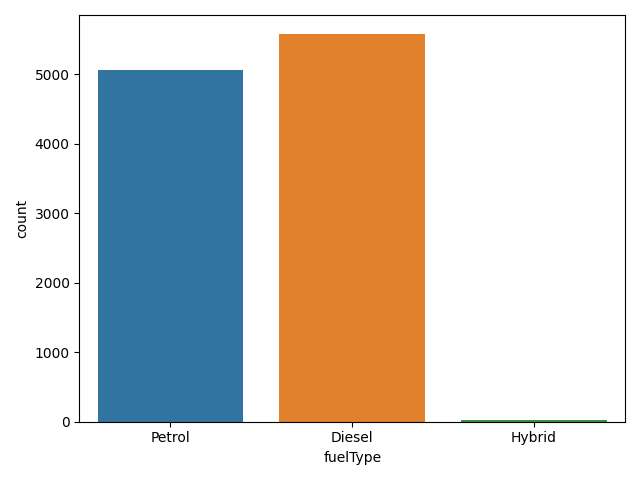
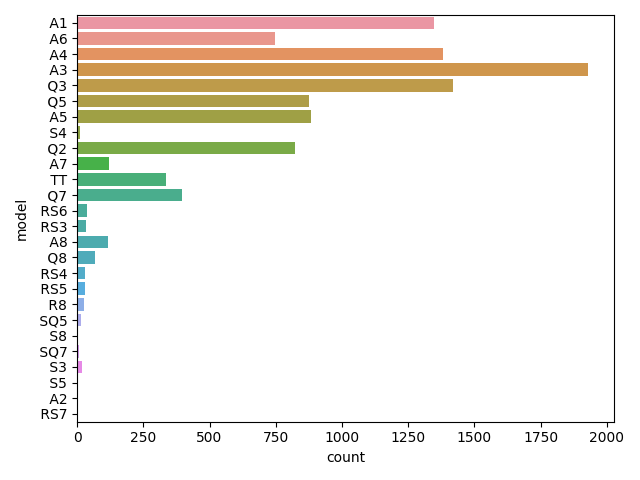
The third algorithm is Random Forest Regression. Random forest regression utilizes the idea of ensemble learning, which is a technique that can take advantage from multiple machine learning algorithms such that it can produce a more accurate prediction.

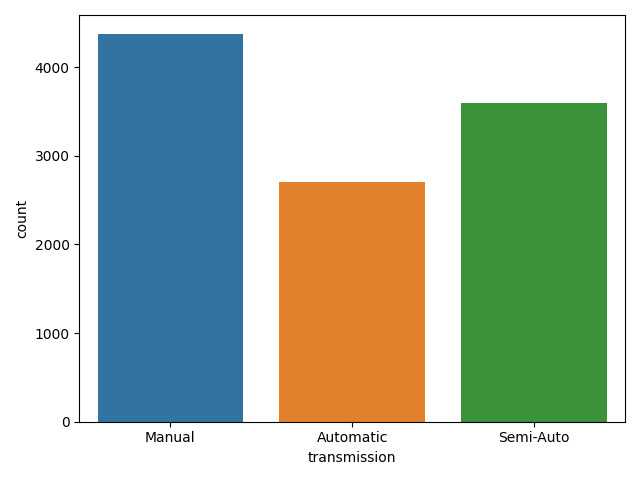
**Detailed Procedures**

1. Exploratory Data Analysis

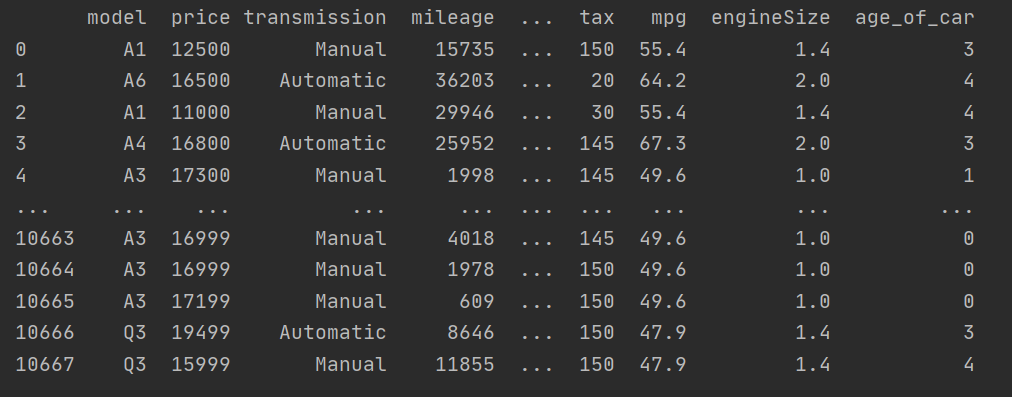


After importing the data, I noticed model, transmission, and fuelType has string values. Then, I used countplot to show the counts of each type of observation in each attribute.





1. Preprocessing



Firstly, compute a new attribute called “age\_of\_car” by subtracting 2020 from the ‘year’ attribute. The result is number of years that car has been used since its registration. I believe the age of car will be a more informative and directly related attribute than its original.

In addition, we need to encode categorical values into numerical values. There are various ways to encode data. The encoding selected is one-hot encoding. Moreover, separate the complete table into X and Y set. X set is independent variables containing all the attributes except ‘price’. Y set is the target variable containing just the ‘price’ attribute. Lastly, apply standardscaler to X set to standardize the data. At this stage, we can start split the training set and test set. By using train\_test\_split helper function, we can split X and Y into X\_train, X\_test, Y\_train, Y\_test.

1. Modeling
2. Accuracy & Evaluation

Different error metrics are required to reveal how well the models fit the training set. Firstly, I calculated the mean absolute error to measure the average discrepancy between predicted values and the actual values. To change the perspective, I calculated the mean absolute percentage error to measure the average error to the percentage of the actual values.

5-fold cross validation is used to evaluate how well the model generalize on unseen data. This can be done readily by applying cross\_val\_score function of sklearn. The training set is split into 5 subsets and in each cross-validation iteration, one non-repeatable subset is selected as validation and the other 4 subsets is used for training. The error in each iteration is stored such that average error across 5 folds can be calculated.

**Comparison of Results**