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 been continuously growing in recent years. However, many factors make the prices fluctuate. For instance, cars have diverse conditions and the market trend is not stationary all the time. 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 on 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>

These datasets contain scraped data of used cars in the British market and are separated into files specific for each car manufacturer. In this report, the dataset selected is “Audi.csv”, that is the file 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.

**Algorithms Overview**

The first algorithm I used is multivariate linear regression. In this hypothesis, multiple independent variables may influence the dependent variable. In the context of our dataset, multiple factors are affecting the price of cars and there is a potential linear relationship between them, which makes it a straight-forward and promising model.

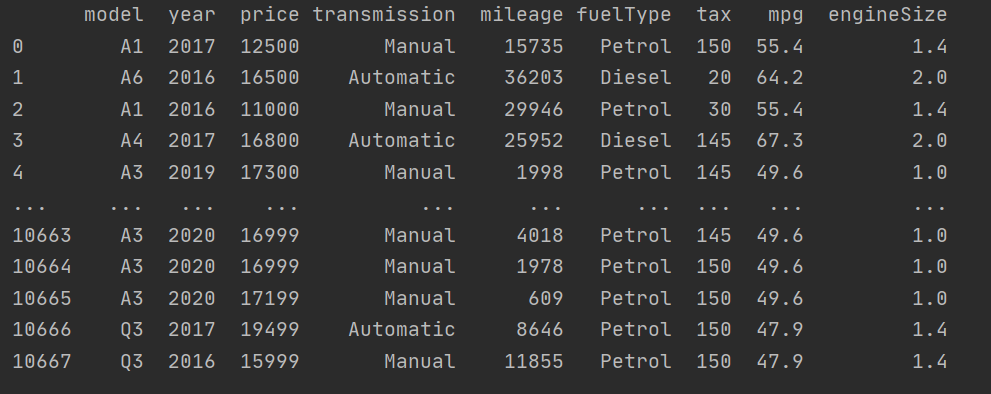
The second algorithm is support vector regression. Support vector regression adheres to the basic principle of support vector machine, which is the maximum margin characteristic, but it is used for regression instead. Linear kernel is used in the SVR.

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

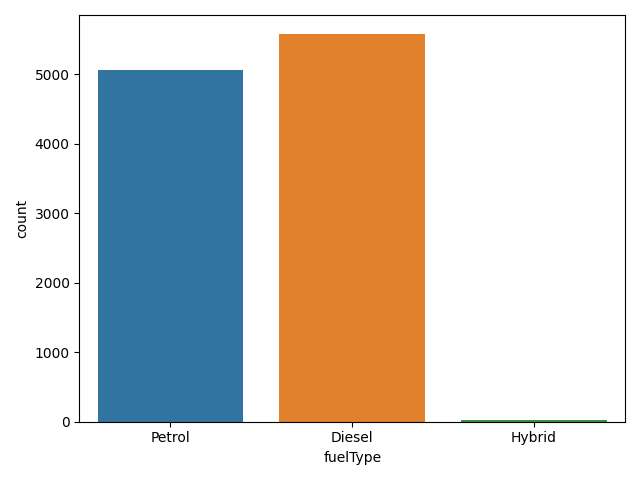
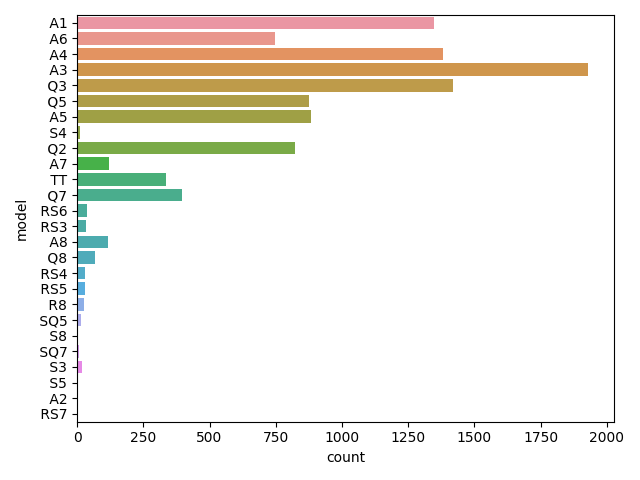
The python package used in this report to train models is sklearn. Before applying the above-mentioned algorithms, it is necessary to normalize our dataset as it is a common requirement of sklearn and it is also considered good practice. The normalization technique used is standardization, which will make the data have zero mean and unit variance.

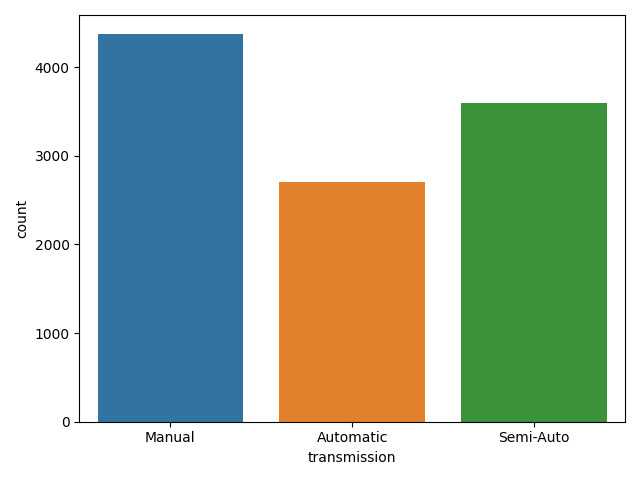
**Detailed Procedures**

1. Exploratory Data Analysis



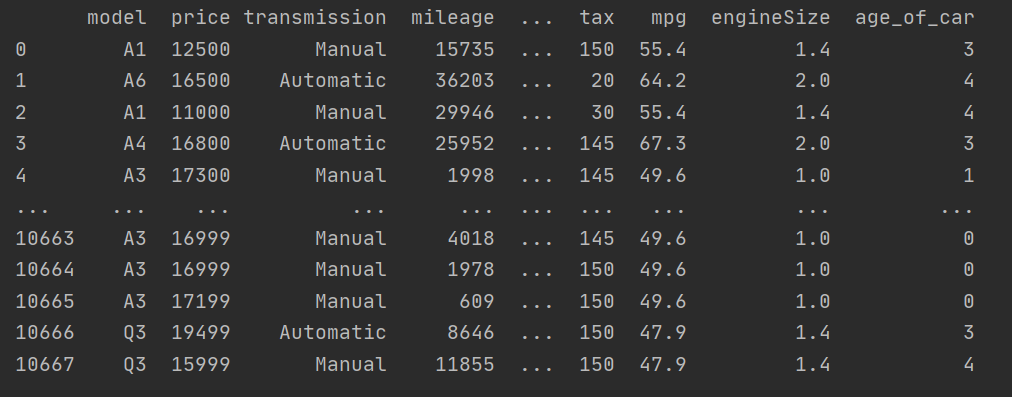
Clearly, model, transmission, and fuelType has nominal values. I used countplot function to plot the counts of each type of observation versus that attribute.





We can see there are finite number of possibilities in each attribute’s value, so we are going to use one-hot encoding.

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, use one-hot encoding to transform categorical values into numerical values. Moreover, separate the complete table into X (features) and Y (target) set. X set is independent variables including all the attributes except ‘price’. Y set is the target variable --‘price’ attribute. Lastly, apply standardscaler on X set to standardize the data. At this stage, we can split the training set and test set. By using train\_test\_split helper function, we are able to split X and Y into X\_train, X\_test, Y\_train, Y\_test.

1. Modeling

By using sklearn, we can easily create regression model to fit X\_train and Y\_train data.

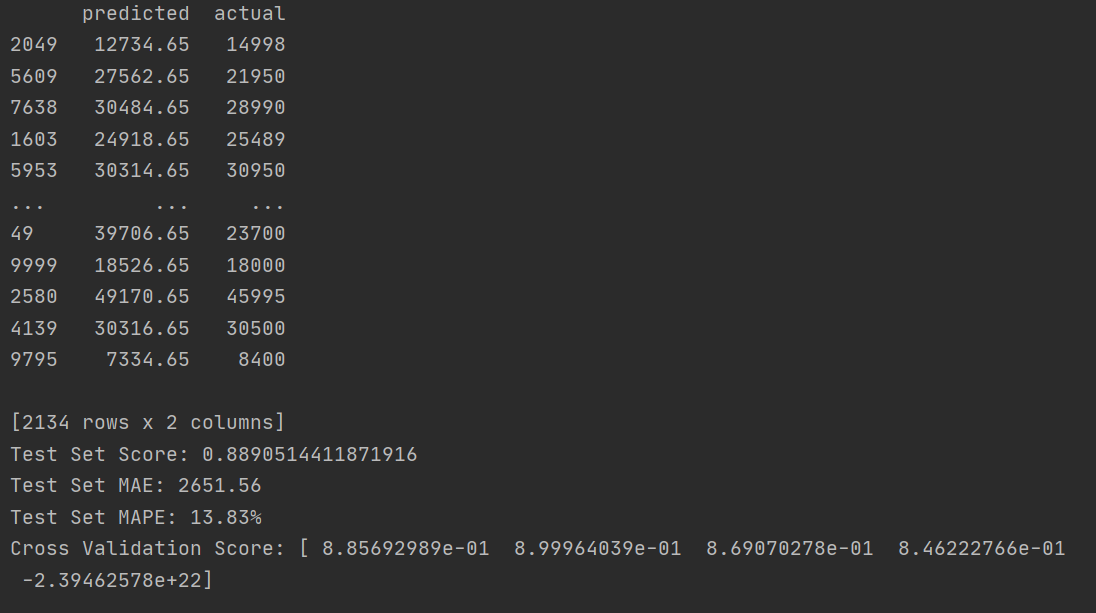
1. Accuracy & Evaluation

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.

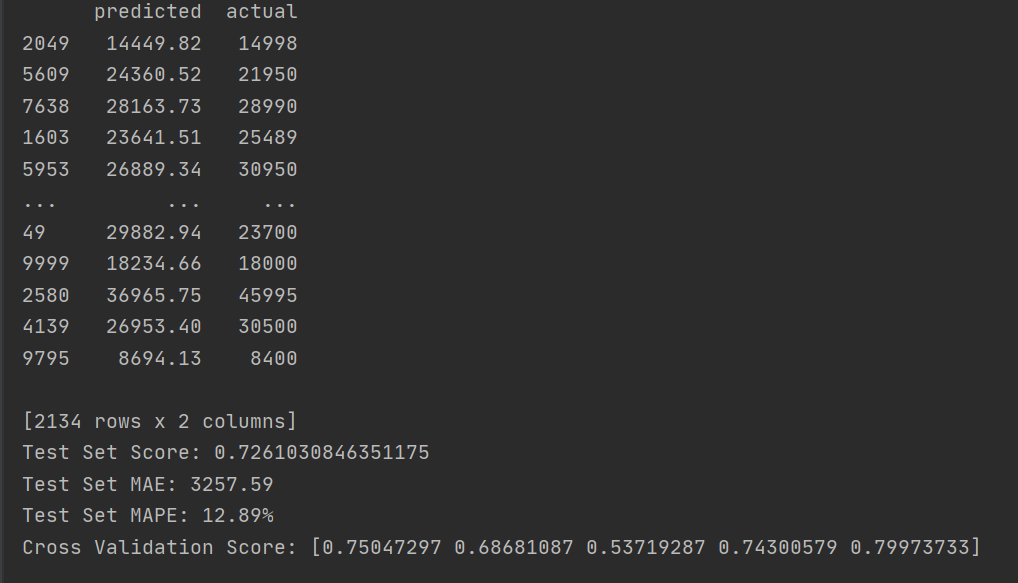
5-fold cross validation is used to evaluate how well the model generalize on unseen data. This can be done readily by applying cross\_val\_score function of sklearn. The training set is split into 5 subsets and in each cross-validation iteration, one non-repeatable subset is selected as validation and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

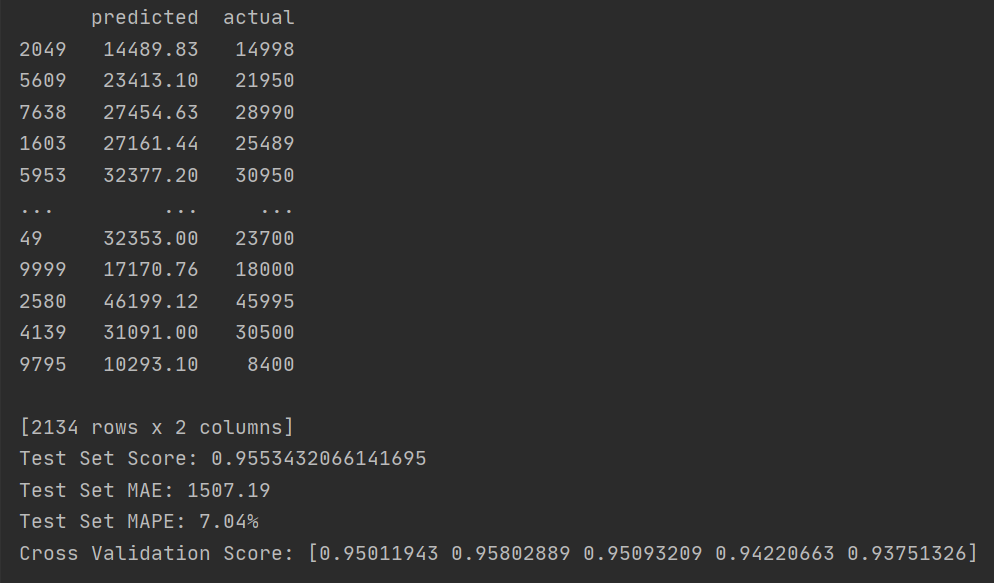
Multivariate Linear Regression



Support Vector Regression



Random Forest Regression



We can see random forest regression has the highest R^2 score (coefficient of determination，best possible is 1.0) in the test set and in 5-fold cross validation. Correspondingly, it has the lowest mean absolute percentage error (MAPE) of 7.04%, which indicates that the average error is 7.04% to the actual values.

Support vector regression has a slightly better MAPE than multivariate linear regression. However, it has a much lower R^2 score which means the support vector regression may perform better on this dataset but may not generalize well on unseen samples.

Multivariate linear regression has a MAPE of 13.8%, which could be acceptable depends on the tolerance of customers. Generally, it has a 0.88 R^2 score which means it is qualified to be used in future new samples. It is worth noticing that in the 5th fold cross validation, the model has a negative R^2 score. This is because the model can be arbitrarily worse as explained by the sklearn documentation.