ECE 9063 Data Analytics Foundations

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 decent conditions and with fair prices. It is also the reason that the market has been growing in recent years. However, it is difficult to choose the opportune moment to buy or sell as the price fluctuates constantly . And there are many factors contributing to the price fluctuations. 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. With the help of a suitable model, buyers will be able to make sure the car is worthy of its price, and sellers can get a more accurate price estimation in accordance with other cars having similar conditions. 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 information.

Dataset Description

Link to the data: https://www.kaggle.com/adityadesai13/used-car-dataset-ford-and-mercedes

These datasets list 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”. It contains 9 attributes and 10668 samples. The dataset is suitable for this assignment as it has adequate attributes and samples. With over 10,000 samples, it is easier to strike a balance between computational time and reliability of the model . 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 so far
- 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 fuelType have nominal 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 while estimating the price of cars in the real-world.

Algorithms Overview

The first algorithm is support vector regression (SVR). 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. Since there are multiple independent variables having potential linear relationship with the target variable, linear kernel is used.

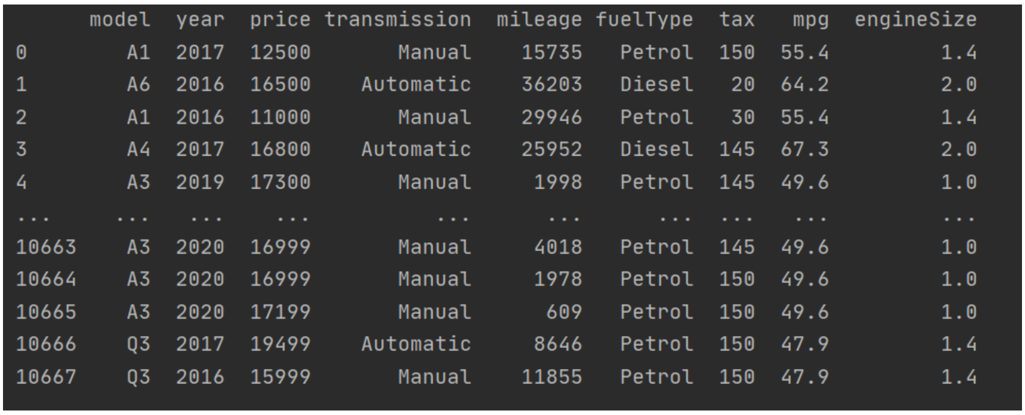
The second algorithm is decision tree regression. The algorithm proceeds incrementally as breaking down the data into smaller subsets and build the associated sub-trees from them. At the end, a tree structure with decision nodes and leaf nodes is constructed. However, one major issue with decision tree regression is that it is very prone to overfitting. At the result comparison section, we will inspect whether this problem arises.

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.

Before applying the above-mentioned algorithms, it is necessary to normalize our dataset as it is a common requirement of many machine learning algorithms 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

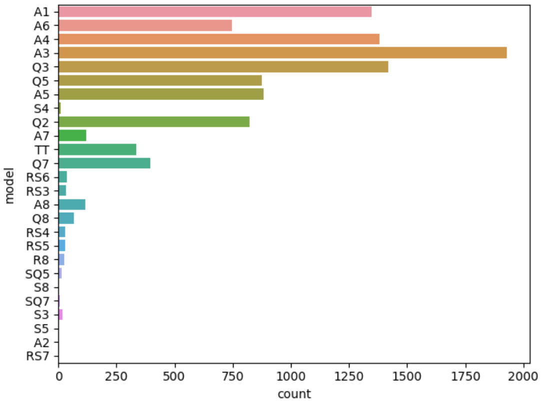
1. Exploratory Data Analysis

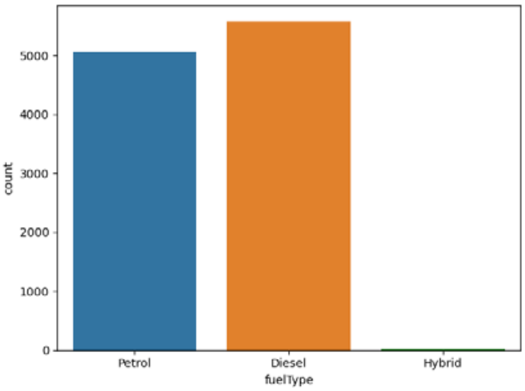


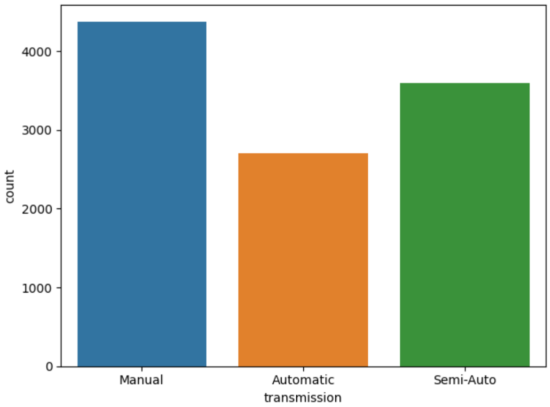
```
      model  year  price transmission  mileage fuelType  tax   mpg  engineSize
0        A1  2017  12500       Manual    15735   Petrol  150  55.4         1.4
1        A6  2016  16500    Automatic    36203   Diesel   20  64.2         2.0
2        A1  2016  11000       Manual    29946   Petrol   30  55.4         1.4
3        A4  2017  16800    Automatic    25952   Diesel  145  67.3         2.0
4        A3  2019  17300       Manual     1998   Petrol  145  49.6         1.0
...     ...   ...    ...          ...      ...      ...  ...   ...         ...
10663    A3  2020  16999       Manual     4018   Petrol  145  49.6         1.0
10664    A3  2020  16999       Manual     1978   Petrol  150  49.6         1.0
10665    A3  2020  17199       Manual      609   Petrol  150  49.6         1.0
10666    Q3  2017  19499    Automatic     8646   Petrol  150  47.9         1.4
10667    Q3  2016  15999       Manual    11855   Petrol  150  47.9         1.4
```

After importing the csv file, I printed the data to have an overview of its structure.

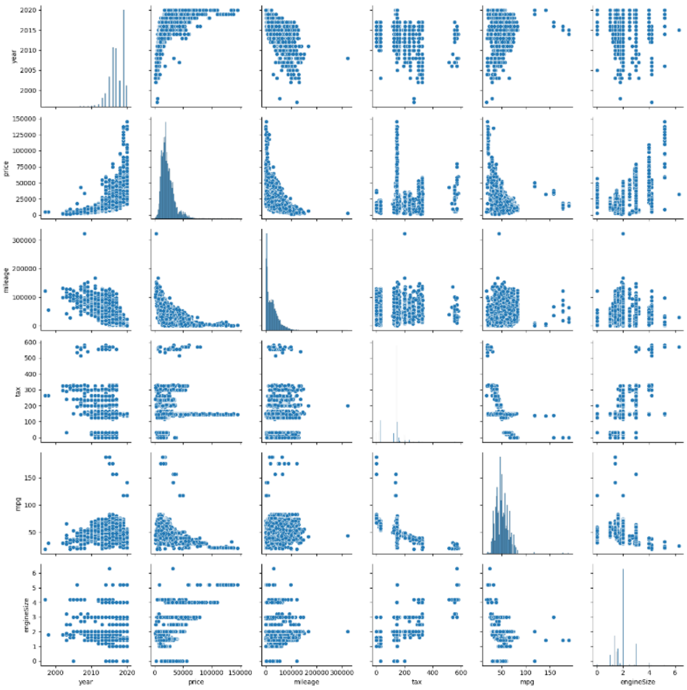
Clearly, model, transmission, and fuelType has nominal values. I plotted the counts of each type of observation versus that attribute.





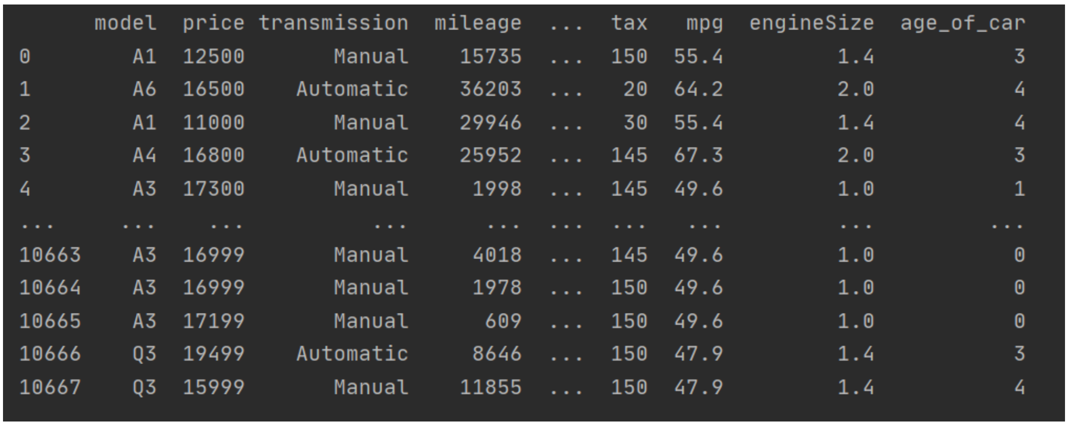


It is observed that there are finite number of possibilities in each attribute, so one-hot encoding is selected as encoding scheme.



From the pair plot graph, we can see the data is dense (No zeros). Therefore, standardization will be used to normalize the data.

2. Preprocessing



```
      model  price transmission  mileage  ...  tax   mpg  engineSize  age_of_car
0        A1  12500       Manual    15735  ...  150  55.4         1.4           3
1        A6  16500    Automatic    36203  ...   20  64.2         2.0           4
2        A1  11000       Manual    29946  ...   30  55.4         1.4           4
3        A4  16800    Automatic    25952  ...  145  67.3         2.0           3
4        A3  17300       Manual     1998  ...  145  49.6         1.0           1
...     ...    ...          ...      ...  ...  ...   ...         ...         ...
10663    A3  16999       Manual     4018  ...  145  49.6         1.0           0
10664    A3  16999       Manual     1978  ...  150  49.6         1.0           0
10665    A3  17199       Manual      609  ...  150  49.6         1.0           0
10666    Q3  19499    Automatic     8646  ...  150  47.9         1.4           3
10667    Q3  15999       Manual    11855  ...  150  47.9         1.4           4
```

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 features into numerical features. Moreover, separate the dataset 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 standard scaling (Standardization) on X set to normalize the input data. At this stage, we can split the normalized features and target into X_train, X_test, Y_train, and Y_test. The test set size is 20% of the entire data.

3. Modeling

By using sklearn, we can easily create regression model to fit the training data. When creating the regression model instance, all the optional arguments are set to default to keep simplicity. All three algorithms will expect two arguments to pass into the function, one being X_train and the other being Y_train. This corresponds to the idea of train the model on the training set. After fitting the model, we can predict the corresponding value of X_test by applying the fitted parameters, then compare the predicted results with the actual target values, Y_test.

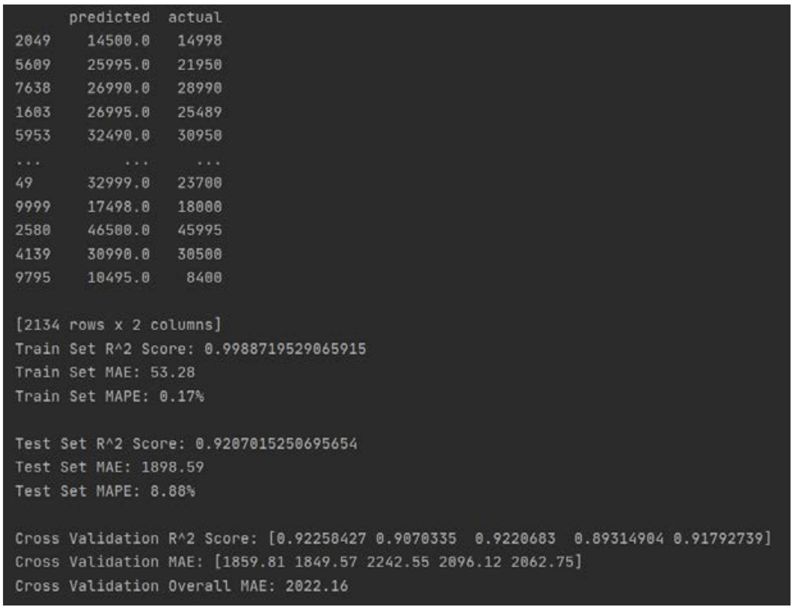
4. Accuracy & Evaluation

Different metrics are applied to reveal how well the models fit the data. Firstly, we will compute the R^2 score, which is the coefficient of determination, as an indication of goodness of fit. It represents the proportion of variance in the dependent variable (Y) that has been explained by the independent variables (X's) in the model. Therefore, it is also a measure of how well the model will perform on unseen data.

Then, I used the mean absolute error to measure the average discrepancy between predicted values and the actual values. To change the perspective, I also used the mean absolute percentage error to show the error as percentage values of the target. Both training errors and test errors are reported. Generally, it is anticipated that the test error will be larger than the training error. We will use test errors as a measurement of how well the model fits.

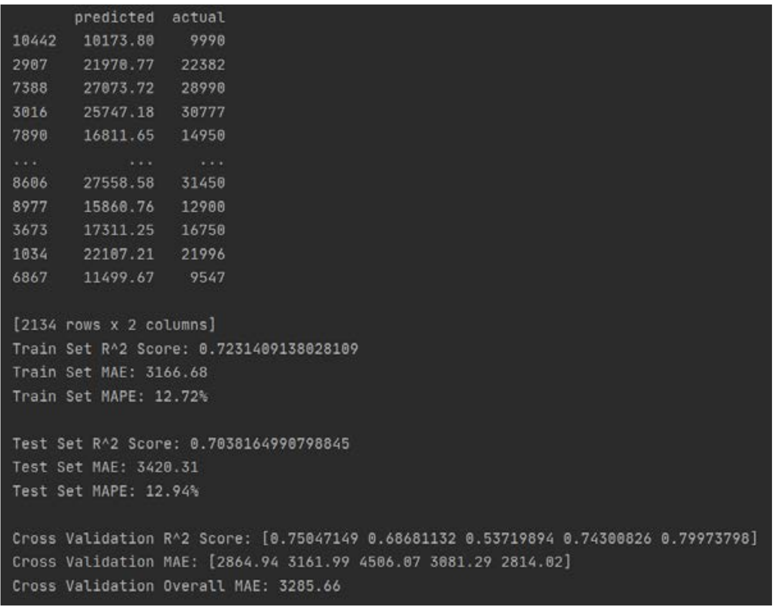
5-fold cross validation is also used to evaluate how well the model generalize on new data. The entire dataset is split into 5 subsets and in each cross-validation iteration, one non-repeatable subset is selected as validation and the other 4 subsets is used for training. The validation error in each iteration is stored to allow computation of overall error for the 5 folds by simply taking average value.

Comparison of Results



```
      predicted  actual
2049    14500.0   14998
5609    25995.0   21950
7638    26990.0   28990
1603    26995.0   25489
5953    32490.0   30950
...         ...     ...
49      32999.0   23700
9999    17498.0   18000
2580    46500.0   45995
4139    30990.0   30500
9795    10495.0    8400

[2134 rows x 2 columns]
Train Set R^2 Score: 0.9988719529065915
Train Set MAE: 53.28
Train Set MAPE: 0.17%

Test Set R^2 Score: 0.9207015250695654
Test Set MAE: 1898.59
Test Set MAPE: 8.88%

Cross Validation R^2 Score: [0.92258427 0.9070335  0.9220683  0.89314904 0.91792739]
Cross Validation MAE: [1859.81 1849.57 2242.55 2096.12 2062.75]
Cross Validation Overall MAE: 2022.16
```

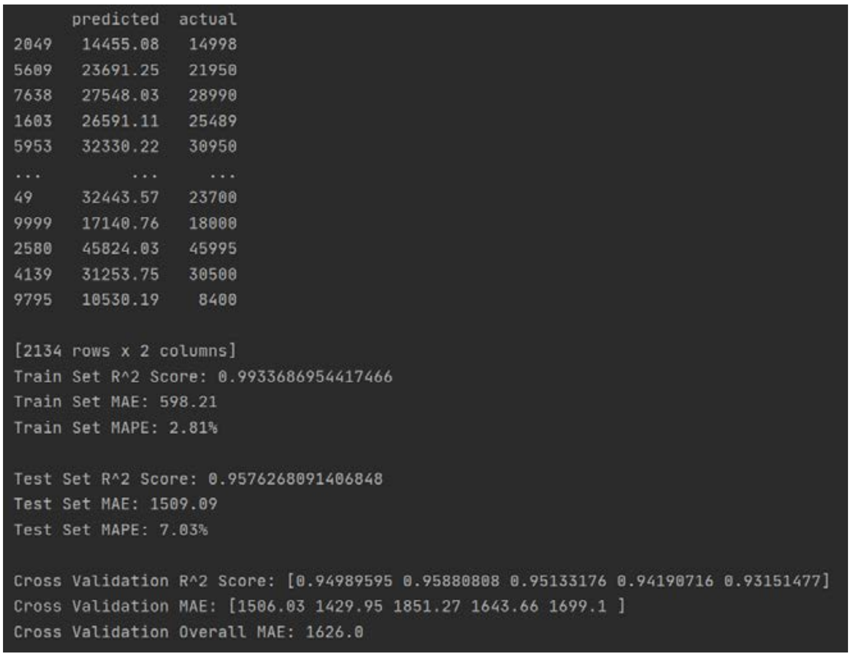
Decision Tree Regression



```
       predicted  actual
10442   10173.80    9990
2907    21970.77   22382
7388    27073.72   28990
3016    25747.18   30777
7890    16811.65   14950
...          ...     ...
8606    27558.58   31450
8977    15860.76   12900
3673    17311.25   16750
1034    22107.21   21996
6867    11499.67    9547

[2134 rows x 2 columns]
Train Set R^2 Score: 0.7231409138028109
Train Set MAE: 3166.68
Train Set MAPE: 12.72%

Test Set R^2 Score: 0.7038164990798845
Test Set MAE: 3420.31
Test Set MAPE: 12.94%

Cross Validation R^2 Score: [0.75047149 0.68681132 0.53719894 0.74300826 0.79973798]
Cross Validation MAE: [2864.94 3161.99 4506.07 3081.29 2814.02]
Cross Validation Overall MAE: 3285.66
```

Support Vector Regression

Random Forest Regression



```
      predicted  actual
2049   14455.08   14998
5609   23691.25   21950
7638   27548.03   28990
1603   26591.11   25489
5953   32330.22   30950
...         ...     ...
49     32443.57   23700
9999   17140.76   18000
2580   45824.03   45995
4139   31253.75   30500
9795   10530.19    8400

[2134 rows x 2 columns]
Train Set R^2 Score: 0.9933686954417466
Train Set MAE: 598.21
Train Set MAPE: 2.81%

Test Set R^2 Score: 0.9576268091406848
Test Set MAE: 1509.09
Test Set MAPE: 7.03%

Cross Validation R^2 Score: [0.94989595 0.95880808 0.95133176 0.94190716 0.93151477]
Cross Validation MAE: [1506.03 1429.95 1851.27 1643.66 1699.1 ]
Cross Validation Overall MAE: 1626.0
```

We can see random forest regression has the highest R^2 score in the test set, also in 5-fold cross validation. Correspondingly, it has the lowest mean absolute error (MAE) and mean absolute percentage error (MAPE) of 7.03%, which indicates the average error is 7.03% of the actual values. Random forest regression also has the lowest overall MAE in cross validation.

Support vector regression has the highest MAE and MAPE in the test set. What is worse still, it has a much lower R^2 score which means the support vector regression may not generalize as good as the other two models in new data.

Decision tree regression has a nearly 1.0 R^2 score, stunningly low MAE (53) and MAPE (0.17%) in the training data. Those are the evidences of overfitting, which means the model fits the training data too well and it may easily fail to capture systematic pattern in unseen data. In the test set and cross validation set, the error becomes larger and close to results of other two regression models.

As anticipated, the MAE and MAPE of the test set are larger than those of the training set in all three algorithms. And the R^2 scores of test set are lower than the training set because the model is not fitted to the test data, it may not perform as good as it was in the training stage.

To sum up, random forest regression has the overall best performance in the three models. And its cross-validation performance is stable, which makes it a suitable model on this dataset and our forecasting problem. Support vector regression is not performing well as it has relatively large errors. And decision tree regression suffers heavily from overfitting, thus it is not recommended on this problem.