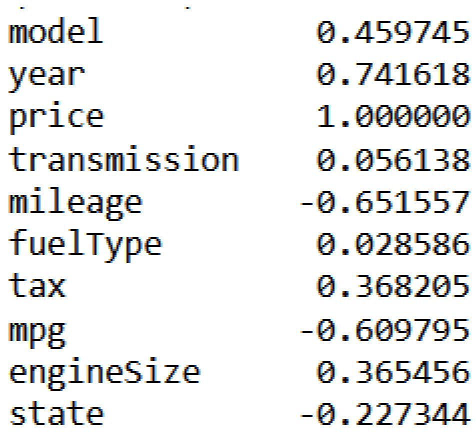
Big Data Analytics – Assignment

Nadin Magdy - 1170453 Ahmed Hany – 1170428, rooms 3708 and 3707 respectively

Steps of solution:

1. Cleaning and Preparing Data

- We load our data and start to see if any cleaning is needed
- We want to look for missing data, and so think how we can fill them in
- When checking for data and any missing information we find:
  - There are 10 missing values for state
  - There are 7 missing mileage values
  - There are 8 models missing
- The numbers of rows with missing data is small, so we have 2 options: either calculate their values or drop these rows. We chose to drop these rows, but we chose to fill in these missing value.
  - The state can be predicted using the mileage, if the mileage is more than 0 we can set the state as used.
  - The mileage can be set according to state, if new just set the mileage to 0. Otherwise we can use the mean of the mileages. Using the average would wrongly skew the data, so we used the median. We could've taken average without the zeroes also.
  - The model cannot be calculated, so we decide an assumption for its missing values to put the most repeated/frequent model in the dataset, which turns out to be A3.
- In models, we appear to have 2 models that we assume are anomalies, model A1. Is the same as A1 and A5. Is the same as A5
- The mpg appeared to have outliers, a negative value, so we replaced it with the median of the mpg
- Obviously ID and ownerName will not help in prediction we drop them
- Initial prediction is that there are some linear relations, for instance between as year increases, mpg increases, tax decreases
- For all categorical data, we want to use hot encoding for the model to understand it
- We analyze data after describing it, and check the correlations between the features so far:
  - The biggest correlation with price is with year, and it is a positive correlation
  - Next is the mpg, and it is a negative correlation. The bigger the mpg the smaller the price
  - After that is mileage, with a negative correlation
  - Transmission and fuel type seem to have little or no effect on the price
  - Engine size used to almost zero correlation, till we did the encoding in a way that the smallest engine was given the smallest vale 1, and the largest XXLarge engine was given the value 7. This showed a positive correlation which makes sense



```
model           0.459745
year            0.741618
price           1.000000
transmission    0.056138
mileage        -0.651557
fuelType        0.028586
tax             0.368205
mpg            -0.609795
engineSize      0.365456
state          -0.227344
```

- Last thing before we start finding the best classifier, we drop the price column in order to predict it

2. Models and Prediction

- We tried 5 different classifier models that could help us predict prices:
  - I. Polynomial features
  - II. SVR
  - III. Linear regression
  - IV. Decision tree
  - V. Random forest
  - VI. Bayes Classifier
- Then, loop on training data with K range (1:9) → a parameter to a function to choose best k features from a given dataset (SelectKBest)
- On each possible value of k,
  - I. Extract best k features
  - II. Split train data to train and test for validation
  - III. Initialize and train model with the given features
  - IV. Apply cross validation with splits = 5
  - V. Get average score and append to list of scores to each classifier
- The best accuracy came from the random forest classifier with 94%