CAPSTONE FINAL REPORT
USED CAR PRICE ALLOCATION

Introduction

Due to the pandemic, the problem of unemployment has risen uncontrollably thus leading many families to reconsider their expenses and therefore becoming hesitant to engage in long-term purchases. Car companies are one of the many institutions that rely on customers who engage in long-term purchases. The lack of funds that would regularly be accompanied by customers who are interested in buying new cars, have shifted towards car companies who are specialized in used cars as it costs fairly less and is a short-term investment. One set back that customers face when purchasing used cars, is if they are paying more than what they should be paying. It is fairly crucial to know the market price for the vehicles for the buyer, but also for the seller in order to compare prices to other sellers.

There is a need for a system in which the price of a certain used car can be determined by analyzing a variety of features. These features are very crucial in determining the best possible price for the used car in particular.

In this project, I will use a data set of 17965 used Ford vehicles information of 9 different brands.

Problem Statement

This project will build a supervised regression machine learning model to predict the price of a used-Ford model vehicles based on their different features such as year of the car, mileage, fuel type, transmission, etc.

For the regression model, the project will use 5 different machine learning algorithms:

- linear regression,
- random forest,
- decision tree,
- Ridge, and
- Lasso Linear Regression.

Not only the customers those trying to find reasonable, true-prices of vehicles they want, the sellers of the used-car vehicles are also be able to determine true prices of the vehicles according the features of the vehicles they are selling with the prediction model used in this project. The

true values of the vehicles after the models of this project will also lead both sides to allocate their marketing budget and efforts intelligently.

Data Source

Scraped data of used cars listings are the source of this project. Data contains (9) different car manufacturer including Ford.

This project will work on Ford car manufacturer with 17965 entries and 9 different features of these entries.

The data involves:

1 year (year of the vehicle)
2 prices (price of the vehicle-value as pounds)
3 transmission (automatic-manual-semi-automatic)
4 mileage (mileage of the vehicle)
5 Fuel Type (type of fuel like petrol-hybrid-diesel, etc.)
6 Tax (tax value of the vehicle)
7 Mpg (miles per gallon)
8 Engine Size (size of the engine)
9 Model (different models of the Ford like Fiesta, Focus, etc.)

Data Description

In our data, we have 17965 samples and 9 different variables including 3 object type (model, transmission, fuel type), 4 integer type (year, price, mileage, tax), and 2 float type (mpg, engine size) variables. We can also describe the data as 3 categorical and 6 numerical data types.

Data Wrangling

In general, there is not so much problematic entries in the data set. In the ‘year’ of the data set, there is an erroneous coding of 2060 instead of 2016 which is also the mean of the year of the data.

When I run the 'engine Size' column, we observed that there are 51 vehicles recorded with an engine Size of 0.0. This is obviously impossible, and these vehicles must be investigated. We see

that these vehicles account for almost a quarter of a percent of the total dataset (0.28388). As a result, these values are going to be removed.

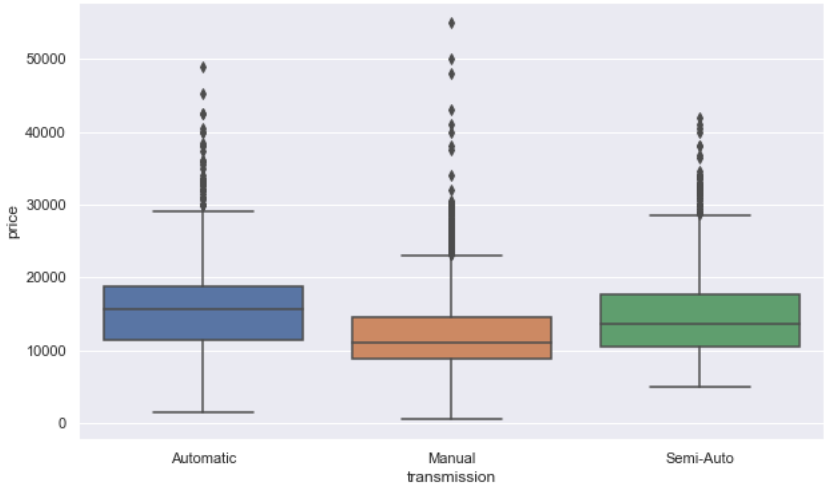
'Mileage' as a variable has also some illogical values which seems to happen while coding process of the data. I will look at the 'mileage' where minimum value=1 that seems to be illogical. I will get deeper to it to understand, which year is equal 1 because, 2020 is OK but, before that year, it is not logical.

As it is seen, there are 6 vehicles with mileage=1 before 2020. This value seems to be erroneous, and the magnitude of these values on overall data is 0.03349. This percentage is not so big, so I can remove value =1 from the vehicles before year 2020.

Exploratory Data Analysis

Categorical Variables

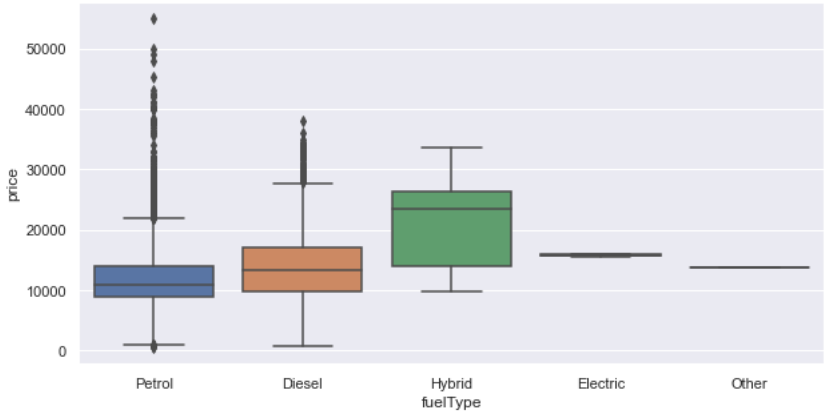
In the first part of the EDA, I want to look at the Categorical variables. As a categorical variable, when we check 'transmission', manual type of vehicles has the highest numbers and there is a significant difference between the others and manual type of vehicles (Manual = 15517, Automatic = 1361, Semi-Auto = 1087).



The vehicles with manual transmission tend to be cheaper to purchase than other transmission types. This may be due to advanced resources required to design and implement automatic

transmission systems. We can clearly see that this feature has a significant influence on the price. In the boxplot, we also see that there are some outliers in each type of vehicles.

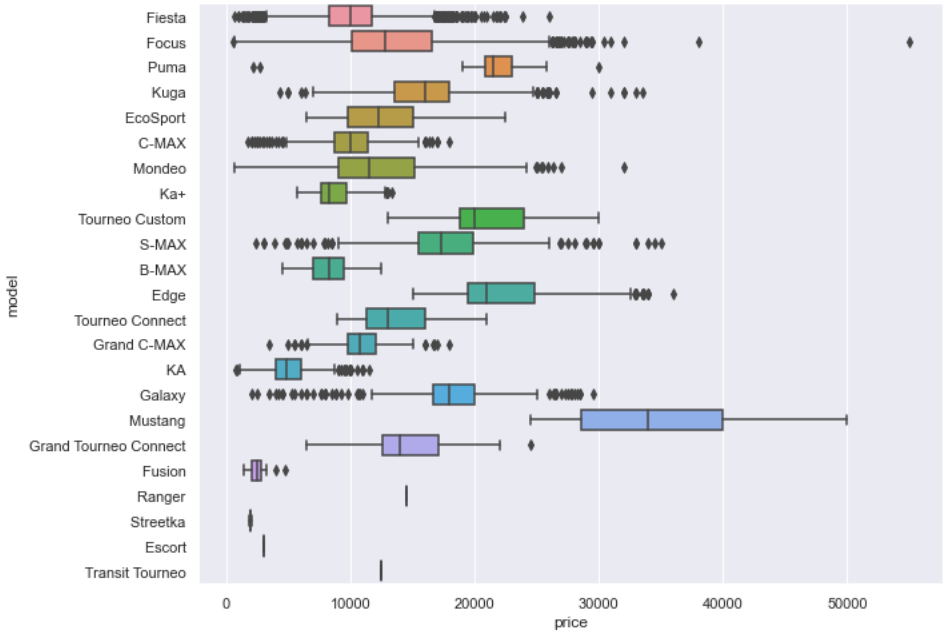
Now, I am going to look at deeper to the ‘fuel Type’ of the vehicles. For vehicles in fuel type usage, vehicles using petrol have significantly highest number (12178) than the others where Diesel (5762) is the second highest number in fuel type vehicles.



We can see that, on average, petrol vehicles are cheaper to purchase than vehicles with different fuel types. Hybrid vehicles are the most expensive to purchase on average, possibly due to the advanced technology required in order to merge petrol and electric motors.

The value counts of electric and other types are very less, and Hybrid vehicles have no outlier values, interestingly. We can clearly observe that the fuel type is an important feature in determining a vehicles sale price.

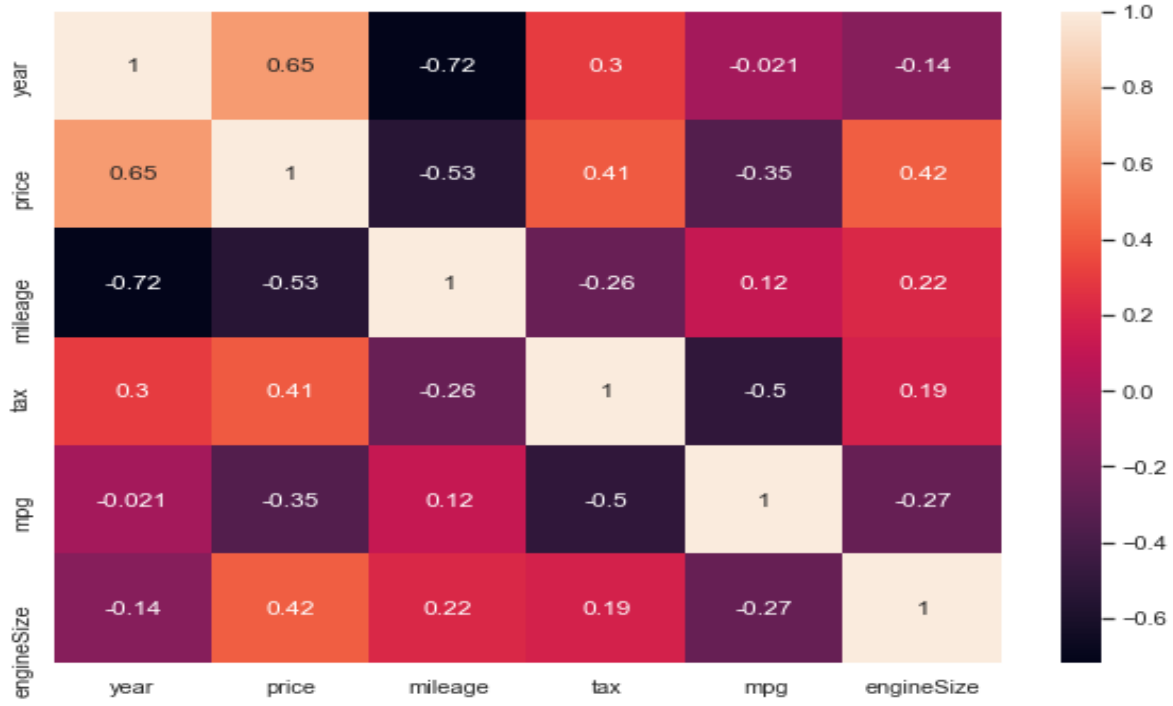
As the last categorical variable, the model of the car has different values where fiesta (6557), focus (4588) model of the Ford vehicles have the highest number when compared to other models.



When we look at to the model of the vehicles, Mustang, Edge, Tourneau and Puma are the most expensive vehicles. There are rare outliers between model and price relationship of the data.

Numerical Variables

In this section we shall analyze the relationships between the numerical variables in our dataset. We shall start by creating a correlation heatmap of the numerical variables.



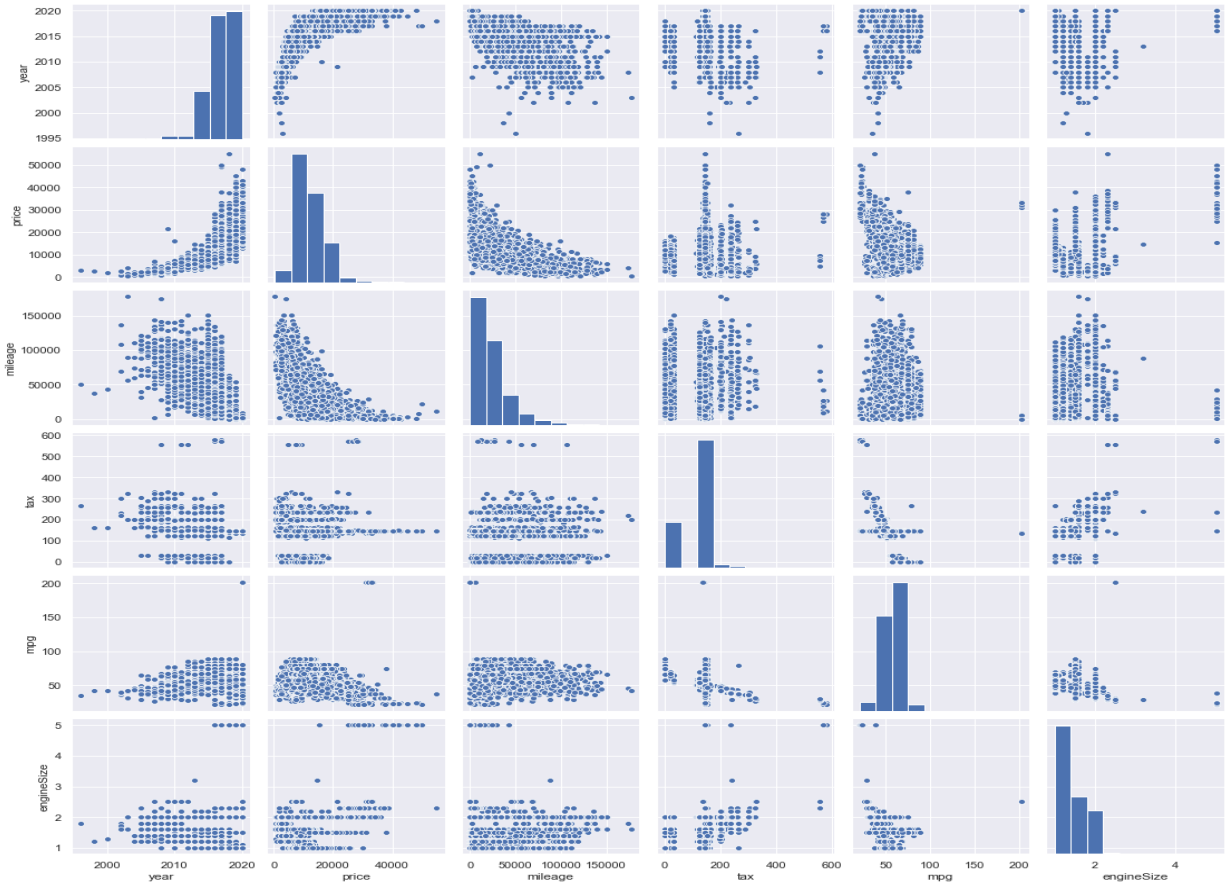
We noticed that there is a positive correlation between year and price and a negative correlation between mileage and price. This makes sense, since newer cars are generally more expensive and cars with more mileage are relatively cheaper. We also notice a negative correlation between mileage and year - the newer a car is the less miles it is likely to have travelled.

Furthermore, we notice a positive correlation between engine size and price, and negative correlation between engine size and mpg. This follows expectation, since it is common practice for manufacturers to sell models with larger engines for a higher price in comparison to the same model with a smaller engine. Also, as the engine size increase, the miles per gallon decreases which is logical.

As a result, due to the higher price, a larger tax payment is required, hence the positive correlation. This also explains the positive correlation between tax and price.

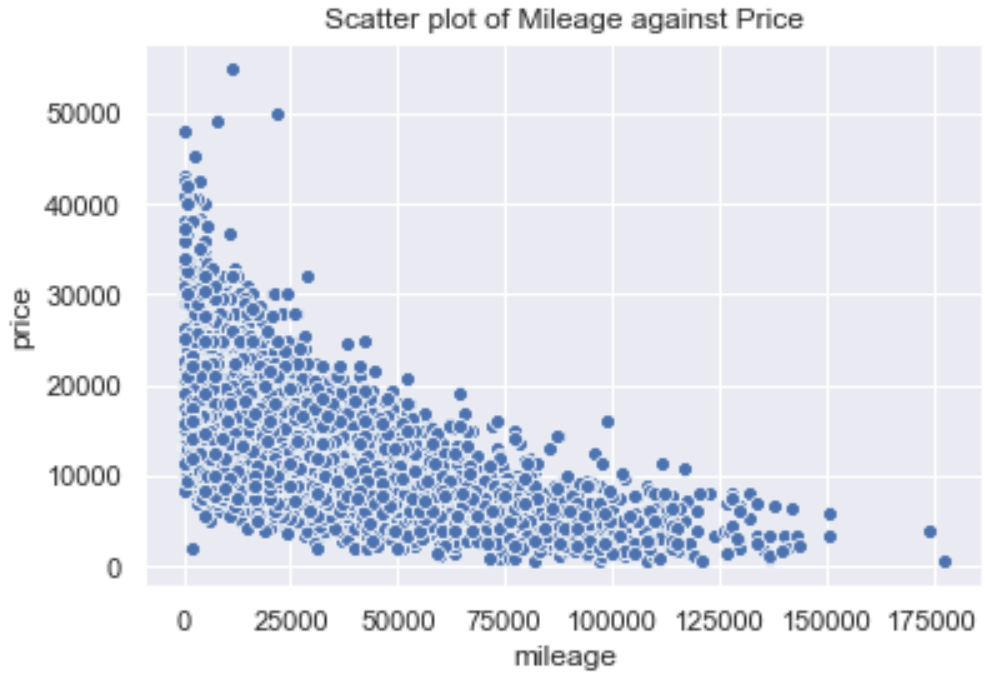
Pair Plot

Let's go deeper and make deeper observations with scatter plots. I will run the pair plot first to get the whole picture of the relationship of the numerical variables.



It is seen that there is relationship between price and other variables either strong or weak.

Now, let's get in deeper and look at the relationship between price and other variables separately.



We notice that the earlier mileage on a vehicle has the most negative impact on the price. This can be seen since the slope on the plot is much steeper for lower mileage, while the rate of decrease of the price reduces as the mileage increases.

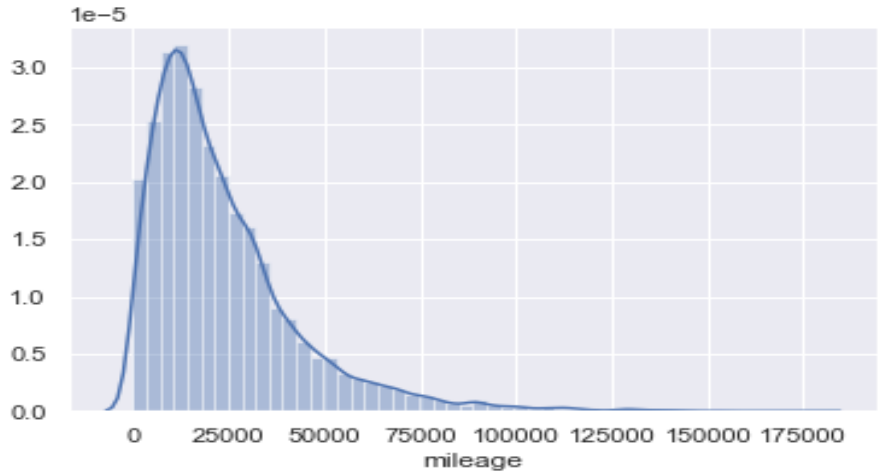
We clearly see that as the engine size of the vehicle increases, the price tends to increase too.

As expected, the new vehicles are much expensive than the older ones.

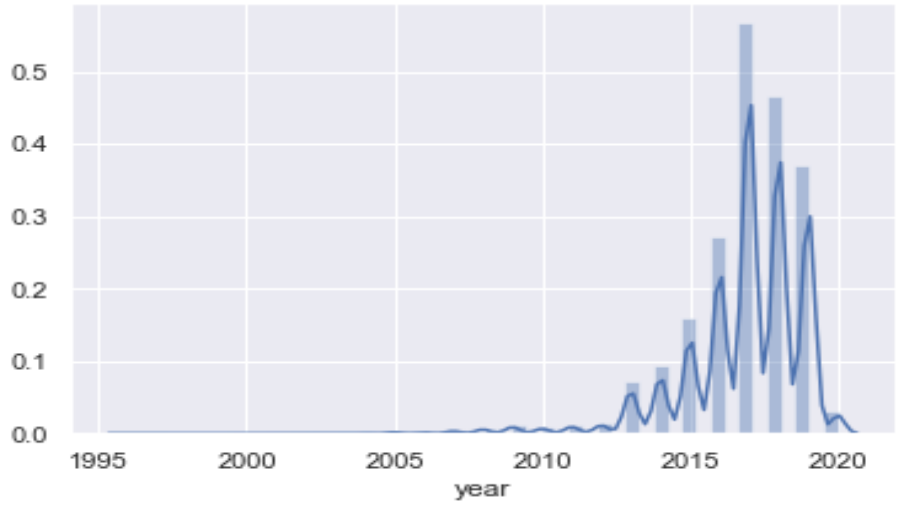
The price range of the vehicles with 25-80 mpg is around 10K and 30K and as the mpg decreases price of the car is also decrease. There is a positive relationship between price and mpg and cheaper cars usually have higher mpg.

Analyzing the Distribution of Numerical Variables

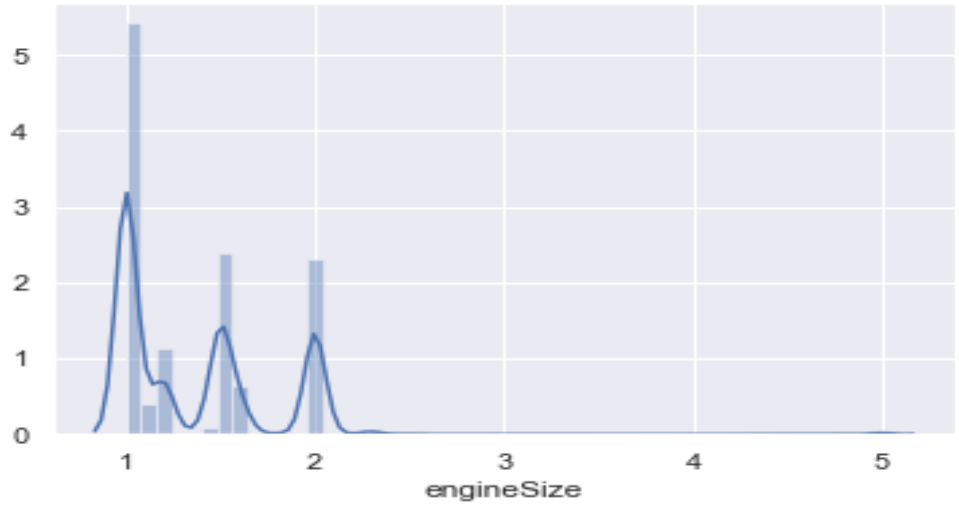
In order to achieve optimized prediction results, we must first ensure that our numerical features are normally distributed. To do this, we produced histograms and check that the follow the "bell" shaped curve.



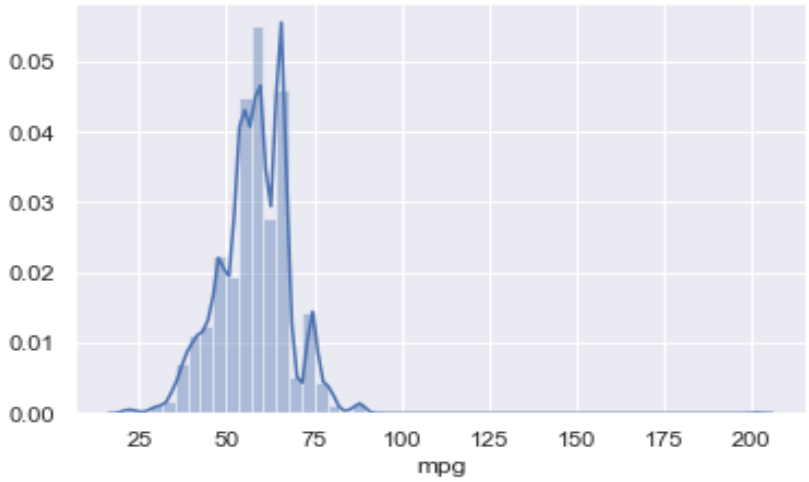
Mileage is positively skewed.



Year is negatively skewed.



Engine Size is positively skewed.



We see a positive skewness in mpg.

Data Processing

Creating Dummy Variables

In order to use our categorical variables in the regression analysis, we must create dummy variables for them. I converted categorical variables into numerical and as next step, I will concatenate these numerical variables with the original data. Also, I removed the original categorical variables from my data set.

After realizing all the steps in data processing, the data will contain 17964 entries and 13 variables. In the data set, it will contain 2 float data type, 4 integer data type, 1 object data type,

and 6 uintegers (primitive data types that are unsigned versions of data which cannot store negative numbers).

INFERENTIAL STATISTICS

Hypothesis Testing

In the inferential statistics part of the project, I will use Automatic type of the vehicle and test if there is any difference on predicting the price of automatic and non-automatic vehicles.

Null = There is no difference between the mean of prices of automatic and non-automatic vehicles.

Alternative = There is difference between automatic and non-automatic vehicles.

I will perform a t-test on two independent samples. I will calculate the value of the test statistic and then its probability (the p-value).

If the p-value is equal or lower than the significance level 0.05, the null hypothesis is going to be rejected.

Hypothesis Testing Result

The p-value in calculations are lower than 0.05 (0.04). It appears that there is a difference between how much is the price of automatic vehicles versus how much non-automatic vehicles prices.

As a result, our hypothesis testing result is we are going to reject null hypothesis and accept the alternative one.

Preprocessing and Training Data Development

Scaling the Data

The difference in ranges of features will cause different step sizes for each feature. To ensure that the gradient descent moves smoothly towards the minima and that the steps for gradient descent are updated at the same rate for all the features, we scale the data before feeding it to the model.
The data must be prepared before building the models. The data preparation process can involve three steps: data selection, data preprocessing and data transformation.
Many machine learning algorithms don't perform as well if the features are not on relatively similar scales.
I will use Standard Scaler to normalize the features of the data. Standard Scaler scales each column to have 0 mean and unit variance.

Creating Training and Test Sets

After scaling the data, there is going to be changes on the values of the columns when we standardize the values of each column.

Next, I split my data into training and test group to understand the model performance those I will use in the model section of my project.

I split dataset by using function train_test_split ().
I will pass 3 parameters; X features without target variable, y as target variable, and test_size. Additionally, I will use random_state to select records randomly.

My target variable is going to be ‘price’ column and assign it as y. Other variables are going to be X.

Modeling

The goal of the modeling step is to develop a final model that effectively predicts the stated goal in the problem identification section. I will review the types of models that would be appropriate for Regression Analysis.

My target variable is going to be continuous ("price"), so I will build the models according to my dataset which will be regression models.

In the modeling part, I will run:

1. Decision Tree Regressor
2. Ridge Regressor
3. Linear Regression
4. Random Forest
5. Lasso Linear Regression

After running the models, I will get the:

1. R-square___; the percentage of the variance in the dependent variable that the independent variables explain collectively. R-squared measures the strength of the relationship between your model and the dependent variable on a convenient 0 – 100% scale.
2. Root Mean Squared Error___; the square root of the variance of the residuals. It indicates the absolute fit of the model to the data–how close the observed data points are to the model's predicted values. RMSE is an absolute measure of fit. As the square root of a variance, RMSE can be interpreted as the standard deviation of the unexplained variance and has the useful property of being in the same units as the response variable. Lower values of RMSE indicate better fit. RMSE is a good measure of how accurately the model predicts the response, and it is the most important criterion for fit if the main purpose of the model is prediction.
3. Average 5-Fold CV Score___; Cross-validation is a powerful preventative measure against overfitting. Cross-validation allows to tune hyperparameters with only original training set. This allows to keep test set as a truly unseen dataset for selecting final model.

Prediction Results

Decision Tree Regressor:
R-square = 0.8957, RMSE = 1513.5, Average 5-Fold CV Score = 0.8789

Our R2 score of 0.89 is good and represents 89% of the variance of the price of a used car based on the independent variables we have used here. Our RMSE value of roughly £1513 is not high and reasonable. Average 5-Fold Cross-Validation score is also reasonable with 0.8789

Ridge Regressor:
R-square = 0.8099, RMSE = 2043.5, Average 5-Fold CV Score = 0.7887

Our R2 score of 0.81 which is lower than the previous model and represents 81% of the variance of the price of a used car based on the independent variables we have used in our dataset. Our RMSE value of roughly £2043 which is also high than previous model. Average 5-Fold Cros-Validation score is lower.

Linear Regression:
R-square = 0.8526, RMSE = 1799.23, Average 5-Fold CV Score = 0.8264

Our R2 score of 0.85 is better than Ridge Model but worse than Decision Tree Model and represents 85% of the variance of the price of a used car based on the independent variables we have used here. Our RMSE value of roughly £1799 is not high and reasonable. Average 5-Fold Cross-Validation score is reasonable.

Random Forest Regressor:
R-square = 0.9224, RMSE = 1240.6, Average 5-Fold CV Score = 0.9224

Our R2 score of 0.93 is good and represents 93% of the variance of the price of a used car based on the independent variables we have used here. Our RMSE value of roughly £1240 is not high and reasonable. Average 5-Fold Cross-Validation score is also good. Random Forest Model is the best model among first 4 models

Lasso Linear Regression:
R-square = 0.8526, RMSE = 1799.2, Average 5-Fold CV Score = 0.8263

Our R2 score of 0.85 is good and represents 85% of the variance of the price of a used car based on the independent variables we have used here. Our RMSE value of roughly £1799 is not high and reasonable. Average 5-Fold Cross-Validation score is also reasonable.

We notice that our scaled random forest regressor is able to explain the most variance (93%) within the price of used cars in comparison to the other models built using the scaled data. We also have the lowest root mean squared error (£1234) as a result of using this model.

Future Improvements

This project used a data with comprised of 100.000 samples from different manufacturers of the vehicles. We only focus on to brand of Ford vehicles and only for England as a country. This project tried to reflect the nature of the used car features and how they can affect the price of it. These predictions may also be occurred in different countries with different models and different features than we have in this data set.