Car Price Prediction Analysis

Data set: [Car Price Prediction Challenge | Kaggle](https://www.kaggle.com/datasets/deepcontractor/car-price-prediction-challenge)

Business Context:

* Car manufacturers set the price of new cars in the industry. With the government adding some additional costs in the form of tax. Due to the ever rising prices in new cars, it can be difficult for people to purchase a car brand new. However, because of the affordability of used cars, people tend to purchase these types of cars instead.
* When either purchasing or selling, it is critical to understand the true market value of a car.
* A used car price prediction system is needed to accurately estimate the cars’ value based on their features (eg mileage, km driven, prod year etc..). While there are already websites that provide this service, their approach may not be as accurate. So several ensemble methods and algorithms may be required in the prediction of a car's true market worth.

.

* As a result, this model established in this study may be of aid to online web services that use a car price prediction system. (TradeMe, 2CheapCars, etc.. ). Many people are interested in the automobile market because they wish to buy or sell
* their car at some point in time.
* It is a mistake to pay too much or undersell your car so it is important to understand the true market value.

Stakeholders: online web services that provide a used car price prediction algorithm

(Car Dealerships)

* TradeMe
* 2CheapCars
* Turners Cars

Business Problem:

* The goal of this project is to estimate the price of cars in the industry, using various machine learning techniques.
* Also for the benefit of buyers and sellers, a predictive analysis on used car prices and estimation on what would be a reasonable price to either buy or sell a car.

Approach:

* Download the data from the website(kaggle) and load into a jupyter notebook. Since this dataset is synthetic there is no web scraping required and can continue to the next step.
* Data cleaning and feature selection would be applied to the dataset first. Once data cleaning is done, I will perform EDA to find patterns in the data and understand how various features are distributed.
* Developing a regression model to predict the price of a car based on various models such as lasso and ridge regression and cross validating the results to find the best possible model for the use of price prediction. This is the case for both the training and testing data.

How the model will be deployed:

* Designing a web page or app to demonstrate the working of the best fit model and provide an easy and interactive way to the user to predict the car based on a few parameters.

Data Cleaning:

* There were no missing values present in the data set
* I have combined the 'Manufacturer' and 'Model' columns together and created a new column called the 'Car Model. I have decided to merge the columns together to reduce the amount of columns in the data set by one. This also puts the vehicle's entire name into a single column rather than split in two, having fewer columns will increase the performance on the machine learning model.
* Created an ‘Age’ column by subtracting the Prod. year by the current year 2022
* Created a Turbo column
* Finding the rows which has turbo in them and assigning the results to the new column Turbo
* Removing the 'ID', 'Manufacturer' and 'Model' now because they are not needed. ID does not provide any information

Data preprocessing:

* There are two characters present, one in levy column with '\_' and one in the Doors column with '>'
* In the 'engine volume' column, some of the values has turbo and some doesn't so we will split and remove the turbo so it will remain the numbers alone
* ‘Mileage’ I am going replace the 'km' text with blank space leaving it as a number, then changing the data type to an integer
* ‘engine volume’ replacing the 'Turbo' text and leaving as a blank speace, then changing the data type to integer
* ‘Levy’ replacing the '-' text to leave as blank, then changing the data type to integer
* ‘Doors’ replacing the '-' abd '>' symbols and changing the data type to integer
* ‘Leather interior’ replacing the yes and no text with 1 and 0
* ‘Turbo’ replacing the true and false text with 1 and 0

Exploratory Data Analysis:

-Split the features into Categorical and Numerical sets

-Created charts for all of the categorical and numerical features to find patterns and trends

Category:

- looking at the category bar graph you can see that the most common vehicle type being used is the Sedan with 8736.The second most common vehicle is Jeep. This makes sense as sedans are more comfortable and appropriate for everyday use by customers.

Fuel type:

- The most commonly used fuel type is petrol(petroleum) as this is widely used by people driving their own private vehicles.Diesel is the second most common type which are mostly used for larger vehicles such as trucks and SUV's.

Gear box type:

- Most of the vehicles produced nowadays are automatic so it is no surprise that most vehicles have an automatic gear box type.

Drive wheels:

- Front wheel drive is the most common with 12874, 4-wheel-drive is not as popular being that they are only used inside the larget trucks and SUV's, amount being 4058

Wheel:

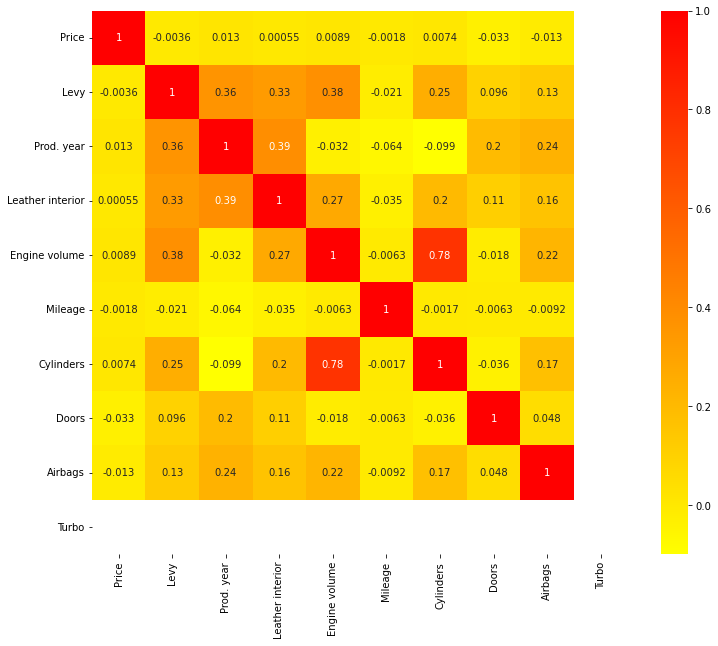
- Since this dataset is based from cars in the US. It is much more common for their steering wheel to be on the left hand side compared to the right hand side like we do here.

Color:

- Black is the most common car amongst all the other colors, White, Silver and Grey are second,third and fourth respectively. People tend to lean more towards buying are dull colored car rather than a vibrant color car.

Numerical Features:

* created a distribution plot for the prod. Year to see how the datas distributed
* There is a gradual increase in production from early 2000's and peaked at around 2014
* The production year of these vehicles range from 1939-2020 however most of the vehicles listed were produced at around 2014-2015 so the price of these vehicles are still relevant considering their age.
* Distribution plot of price suggests that the data is slightly right skewed. This is due to the mean value being larger than the median
* Created a bar chart for the airbags. Most of the vehicles have either 4 airbags or 12 airbags

I then created a heatmap to find any features that may have a strong relationship with the ‘Price’ feature.

There seem to be values that were non sensible.

The minimum price for a vehicle: $1

The maximum price for a vehicle: $26307500

The average price for a vehicle: $18587.435267385332

This required looking into and hopefully removing

Removing Outliers:

I used the IQR (interquartile range) method to remove the outliers that were apparent in the data set.

Q1 = df.quantile(0.25)

Q3 = df.quantile(0.75)

IQR = Q3 - Q1

pos\_outlier = Q3 + 1.5 \* IQR

neg\_outlier = Q1 - 1.5 \* IQR

In doing so, this removed most of the values that were non sensible, either extremely high or too low

Fitting to Model:

For the categorical features I used label encoding to encode the categorical objects to numerical values. This would change the data type of the features from objects to integer or floats.

I understand that OneHot Encoding is the more appropriate approach because label encoding may make it seem as if there are rankings between values. However this did not have much impact on the performance of my model so I used this method instead.

I then split the data in training and testing data sets and used several different models and compared the results for each model.

* Linear regression
* Lasso regression
* Ridge regression
* Gradient Boost regressor
* Ada Boost regressor
* XG Boost regressor
* Random Forest regressor

Each model was put through grid search cross validation to find the best possible parameters for the model. For the boosting models the default base estimators used is decision tree.

After each model was fit and the results were compared. A table and a barchart was made to visualize the different results of all the models. Random Forest model performed the best with 77%

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **R2 Score** | **MAE** | **MSE** |
| Linear | 22% | 7859 | 9973 |
| Lasso | 22% | 7874 | 9968 |
| Ridge | 22.33% | 7859 | 9967 |
| Ada | 42% | 6935 | 8555 |
| Gradient | 75% | 3583 | 5576 |
| XGB | 76% | 3625 | 5533 |
| RF | 77% | 3380 | 5389 |