AUTOMOBILE PRICE PREDICTION USING MACHINE LEARNING

SUBMITTED BY

SWETHA K	1VE17IS054	shweth3@gmail.com
RAKSHITHA G	1BY17CS132	rakshu9676@gmail.com
RISHIKA SANKARAN	1BY17CS140	rishikasankaran@gmail.com

innovation.creation

INSTRUCTOR

ABHISHEK C

Machine Learning Tutor

2020-2021

ACKNOWLEDGEMENT

We, as a team, are extremely grateful to ICS for giving us the wonderful opportunity to pursue this internship. It was a great journey filled with a lot of learning. We extend our special gratitude to Mr. Abhishek, who has been an invaluable source of information and a strong pillar of support. Also, we are grateful to the entire team of members, from ICS, who have aided in the internship journey right from the beginning. Finally, we are grateful to our peers, who have been of great assistance in the attainment of knowledge.

ABSTRACT

There are various factors that determine the price of a car. This makes the price prediction of cars a challenging task. To build a model for predicting the price of cars, we have applied three machine learning techniques (Linear Regression, Decision Trees and Random Forest). We employ these machine learning models that can accurately predict the price of a car based on its features, in order to make informed purchases. We implement and evaluate various learning methods on the dataset. Our results show that the Decision Tree and Random Forest Regressors yield the best results. Conventional linear regression also yielded satisfactory results, although its accuracy was not as high as the aforementioned algorithms.

ABOUT THE COMPANY

ICS is a digital service provider that aims to provide software, designing and marketing solutions to individuals and businesses. ICS believe that service and quality is the key to success. They provide all kinds of technological and designing solutions from Billing Software to Web Designs or any custom demand that we may have.

Some of their services include:

- Development – They develop responsive, functional and super fast websites. They keep user experience in mind while creating websites.
- Mobile Application -They offer a wide range of professional android, iOS & Hybrid app development services for our global clients, from a start-up to a large enterprise.
- Design - They offer professional Graphic design, Brochure design & Logo design. They are experts in crafting visual content to convey the right message to the customers.
- Consultancy – They guide us with expert advice on our design and development requirements.
- Videos - They create polished professional videos.

INDEX

INTRODUCTION

The Python and Machine Learning Internship at ICS, was one with a plethora of learning experiences. A practical knowledge of the Python libraries such as numpy, matplotlib and seaborn, and the various Machine Learning Algorithms was imparted to us. Having worked on Regression and Classification models such as Decision Trees, Support Vector Machines and Logistic Regression, we gained thorough insight into each such concept. Additionally, we learnt about the methods and techniques involved in handling real-world data, covered under the subject of Exploratory Data Analysis, which proved to be invaluable. However, in order to completely absorb the concepts with clarity, and to implement this project, we referred to various websites, including the official documentations of the libraries.

PROBLEM STATEMENT

The prices of new cars in the industry are fixed by the manufacturer with some additional costs incurred by the Government in the form of taxes. There is a need for a car price prediction system to effectively determine the worthiness of the car using a variety of features. Even though there are websites that offers this service, their prediction method may not be the best. Besides, different models and systems may contribute on predicting power for a car’s actual market value. Thus, building a system that accurately determines the price of cars is essential, relevant and necessary.

OBJECTIVE

- The main aim of the project is to give an accurate prediction of the price of a car based on the features describing it.
- To choose 3 machine learning models for training the dataset, to predict the prices with satisfactory accuracy.
- To compare the accuracy of the three models used for prediction.
- To determine the most efficient algorithm among the three chosen models.

REQUIREMENT SPECIFICATION

Hardware Requirements:

- RAM: 512 MB
- Processor: Intel core i3/i5/i7 or any high processor.
- Keyboard
- GPU (preferably) or CPU

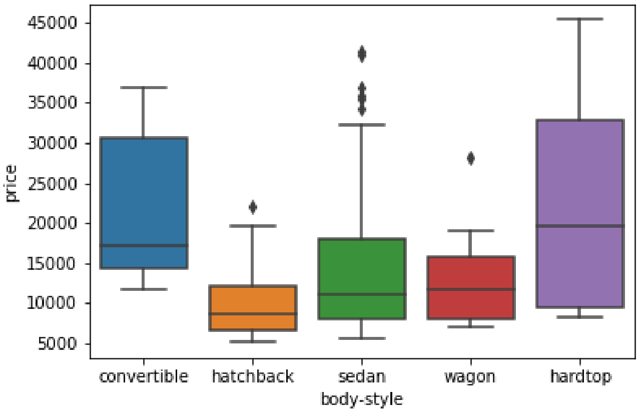
Software Requirements:

- Operating System: Platform Independent
- Application: Anaconda with python libraries installed
- Language: Python 3.6 or above
- Libraries: Numpy, Matplotlib and Seaborn, SciKit-Learn

EXPLORATORY DATA ANALYSIS

1. Price vs Body Style

Plot:



Code:

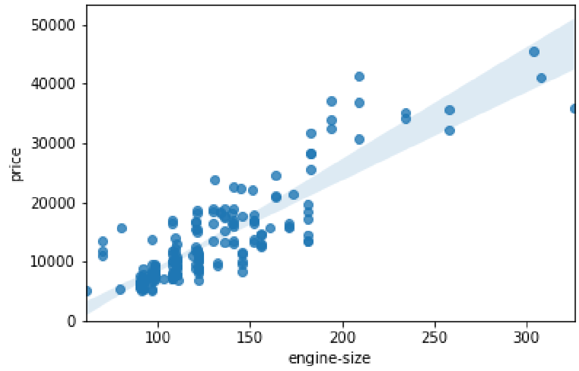
```
sns.boxplot(x="body-style", y="price", data=df)
plt.savefig('2')
```

Analysis:

The distributions of price between the different body-style categories have a significant overlap, and so body-style would not be a good predictor of price.

2. Price vs Engine Size

Plot:



Code:

```
sns.regplot(x="engine-size", y="price", data=df)
plt.ylim(0,)
plt.savefig('1')
```

```
df[["engine-size", "price"]].corr()
```

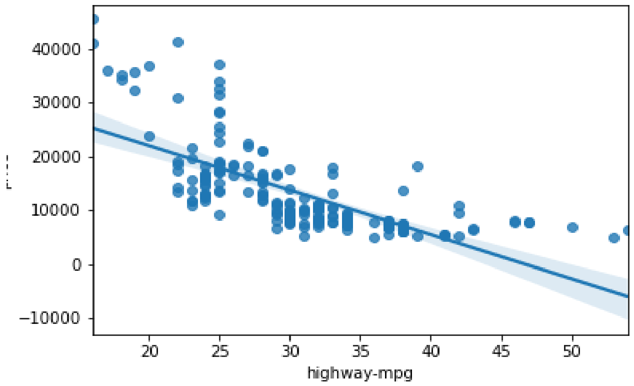
Analysis:

- The above scatterplot of "engine-size" and "price" indicates a positive linear relationship between the two.
- As the engine-size goes up, the price goes up: this indicates a positive direct correlation between these two variables. Engine size seems like a pretty good predictor of price since the regression line is almost a perfect diagonal line.

The correlation between the two is found to be 0.872335.

3. Price versus Highway-mpg

Plot:



Code:

```
sns.regplot(x="highway-mpg", y="price", data=df)
plt.savefig('3')
```

```
df[['highway-mpg', 'price']].corr()
```

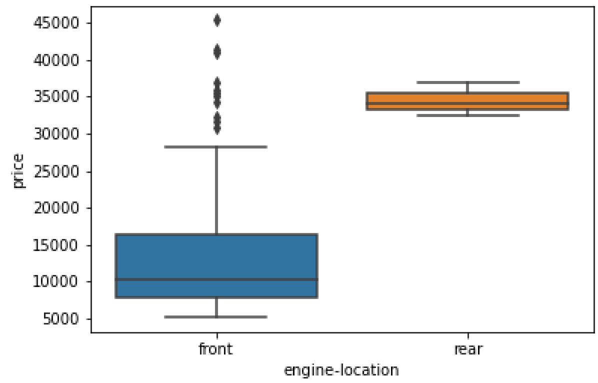
Analysis:

As the highway-mpg goes up, the price goes down: this indicates an inverse/negative relationship between these two variables. Highway mpg could potentially be a predictor of price.

The correlation between the two is found to be -0.704692.

4. Price vs Engine-Location

Plot:



Code:

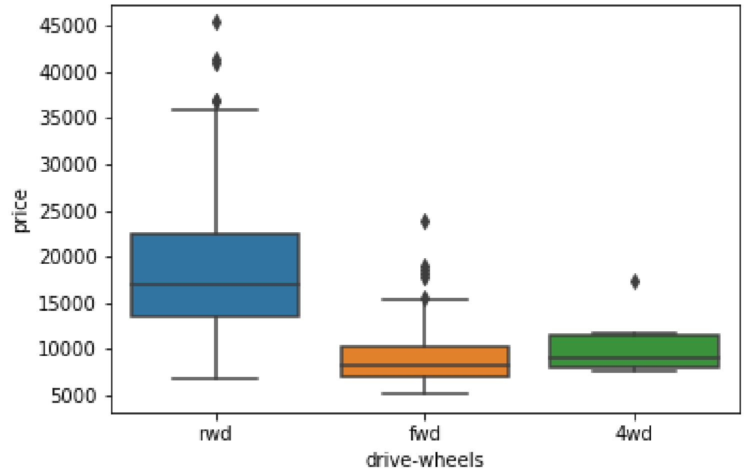
```
sns.boxplot(x=”engine-location”, y=”price”, data=df)
plt.savefig(‘4’)
```

Analysis:

The distribution of price between these two engine-location categories, front and rear, are distinct enough to take engine-location as a potential good predictor of price.

5. Price vs Drive- wheels

Plot:



Code:

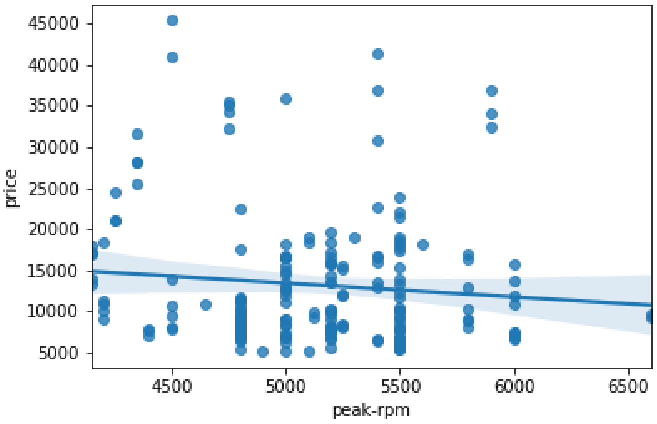
```
sns.boxplot(x="drive-wheels", y="price", data=df)
plt.savefig('5')
```

Analysis:

Here we see that the distribution of price between the different drive-wheels categories differs; as such drive-wheels could potentially be a predictor of price.

6. Price vs Peak-rpm:

Plot:



Code:

```
sns.regplot(x="peak-rpm", y="price", data=df)
plt.savefig('6')
```

```
df[['peak-rpm','price']].corr()
```

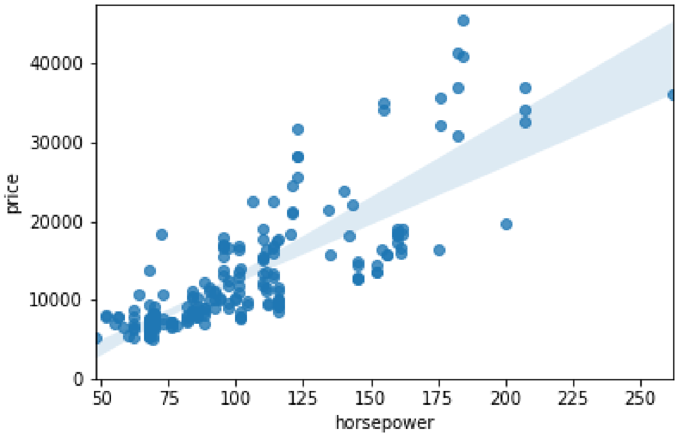
Analysis:

Peak rpm does not seem like a good predictor of the price at all since the regression line is close to horizontal. Also, the data points are very scattered and far from the fitted line, showing lots of variability. Therefore it's it is not a reliable variable.

The correlation between the two is found to be -0.101616.

7. Price vs Horsepower

Plot:



Code:

```
sns.regplot(x="horsepower", y="price", data=df)
plt.ylim(0,)
plt.savefig('7')
```

Analysis:

- As the highway-mpg goes up, the price goes down: this indicates an inverse/negative relationship between these two variables. Highway mpg could potentially be a predictor of price.
- We can examine the correlation between 'highway-mpg' and 'price' and see it's approximately -0.704.

PREPARING MACHINE LEARNING MODELS

- The type of task identified here is regression.
- The following three algorithms are implemented in this project for the given dataset:
  - ✓ Linear Regression
  - ✓ Random Forest (Regressors)
  - ✓ Decision Trees (Regressors)

Linear Regression:

It predicts the outcome of a dependent variable based on the independent variables. The relationship between the variables is linear and hence the word linear regression.

Code:

```
from sklearn.linear_model import LinearRegression
lm = LinearRegression()
lm.fit(X_train,y_train)
print('Coefficients: \n', lm.coef_, '\n')
predictions = lm.predict( X_test)
plt.scatter(y_test,predictions)
plt.xlabel('Y Test')
plt.ylabel('Predicted Y')
print('R2 Score: ', metrics.r2_score(y_test, predictions),'\n')
```

```
R2 Score:  0.8092320637168342
```

Random Forest Regressor:

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap and Aggregation, commonly known as bagging. The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees.

Code:

```
from sklearn.ensemble import RandomForestRegressor

rf = RandomForestRegressor()
rf.fit(X_train, y_train)
predictions=rf.predict(X_test)

plt.scatter(y_test,predictions)
plt.xlabel('Y Test')
plt.ylabel('Predicted Y')


from sklearn import metrics

print('R2 Score:', metrics.r2_score(y_test, predictions))
```

```
R2 Score: 0.9377182890696486
```

Decision Tree Regressor:

Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

Code:

```
from sklearn.tree import DecisionTreeRegressor
regressor = DecisionTreeRegressor(random_state = 0)
regressor.fit(X_train, y_train)
predictions = regressor.predict(X_test)
plt.scatter(y_test, predictions)
print("R2 Score: ", metrics.r2_score(y_test,predictions))
plt.xlabel('Y Test')
plt.ylabel('Predicted Y')
```

```
R2 Score: 0.9012478421695049
```

ML MODEL CHART

Serial Number	Algorithm used	R2 Score
1	Random Forest	0.937718
2	Decision Tree	0.901247
3	Linear Regression	0.8092321

HURDLES

We worked on a real-time dataset, which was quite challenging to analyze. Thus, the exploratory data analysis that we conducted was a cumbersome task. However, there was a lot of learning involved, which gave us exposure to various new concepts. Also, understanding the machine learning models employed in the project was a little hard. Nevertheless, we put in the requisite efforts and overcame the same.

CONCLUSION

Upon exploring the dataset in detail, we found correlations between the attributes and the output. After data cleaning and pre-processing steps, we fed the dataset to three machine learning models that we found to be the most efficient and appropriate for the given task. On observing the results of the predictions, we found all the three algorthims to produce satisfactory results, with two being far more superior to the third, viz, Random Forest and Decision Trees showed much better performace results than the traditional Linear Regession algorithm.

BIBLIOGRAPHY

- https://stackoverflow.com/
- https://www.tutorialspoint.com/index.html
- https://matplotlib.org/3.2.1/api/_as_gen/matplotlib.pyplot.plot.html
- https://www.geeksforgeeks.org/