INFORMATION TECHNOLOGY UNIVERSITY

Artificial Intelligence

Project Report

Car Price Prediction

Submitted By

Muhammad Abdul Rehman (**BSCS20007**)
Muhammad Taqi Raza (**BSCS20031)**

Course Instructor

Mehwish Ghafoor

Abstract

The production of cars has been increasing in the past couple of decades because of increase in demand due to population. According to survey in 2019, over 80 million cars has been manufactured. This has given rise to the cars market and the recent advent of online portal for car purchasing and reviews has facilitated the need for both seller and customer to be better informed about the market trend, prices, etc. Using Machine Learning algorithms like "Linear Regression", we tried to build a model which predicts the car's price based on some available features and information and later predicting the prediction accuracy to check how well our model is working.

Introduction

Car production market is an ever-rising industry. The emergence of online portals such as PakWheels, OLX and many others has assisted the need for both the customer and the seller to be better informed about the trends and price of a car according to its features and model. Machine Learning algorithms can be used to predict the retail value of a car, based on a certain set of features. By training statistical models, we can predict the prices. The main objective of this is to use linear regression based on available data set and compare their levels of accuracy by finding their loss from actual value.

Related Work

Predicting price of a cars has been studied extensively in various research. Listian discussed in her paper, that linear regression model that was built can predict the price of a car that has been leased with better precision than Support Vector Machine. This is on the grounds that Linear Regression (LR) is better in dealing with datasets with already available features and it is less prone to overfitting and underfitting.

Another approach was given by Richardson [1] in his thesis work [1]. According to him, he applied multiple regression analysis to show that hybrid cars retain their value longer than conventional cars. Wu-et-al conducted car price prediction study, by using neuro-network knowledge-based system. He considered the attributes like, Make, year, engine type. Their predictive model gave similar results to a simple regression model. In addition, since car dealers are in high demand to sell cars at the end of the lease year, they have developed an expert system called ODAV (Optimal Distribution of Auction Vehicles). The system provides insight into the best prices on vehicles and where to get the best prices. A regression model based on a linear regression algorithm was used to predict car prices. The system is typically very successful, with over 2 million vehicles replaced.

Gonggie proposed a model built using ANNs [2](Artificial Neural Networks) for used car price prediction. He considered several attributes: Mileage, life expectancy, car brand. The proposed model was built to deal with linear relationships in the data. This was not the case for previous models using simple/SVM techniques. The linear model was able to predict car prices more accurately than other nonlinear models. Furthermore, Pudaruth applied various machine learning algorithms. k-Nearest Neighbors, Multiple Linear Regression, Decision Trees, Naive Bayes for Car Price Prediction in Mauritius. The dataset used to create the predictive model was manually collected from local newspapers over a period of less than a month, as time can have a noticeable effect on car prices.

He used the features like Manufacturer, model, displacement, mileage, year of manufacture, exterior color, type of transmission, price. However, the authors found that naive Bayes and decision trees failed to predict and classify numbers. Also, the limited number of record instances could not provide good classification performance. Accuracy is less than 70%. Noor and Jan [3] use simple linear regression to build a model for automobile price prediction. The dataset was created over the course of two months and included the following features:

Price, engine displacement, exterior color, ad date, ad views, power steering, mileage (km), rim type, transmission type, engine type, city, place of residence, model, version, make, year. After applying feature selection, the authors considered only engine type, price, model year, and model as input features. With the given settings, the author was able to achieve his 98% prediction accuracy. In the above related work, the author proposed a predictive model based on a single machine learning algorithm. However, it is surprising that a single machine learning algorithm approach does not yield outstanding prediction results, and that combining various machine learning techniques into an ensemble can improve it.

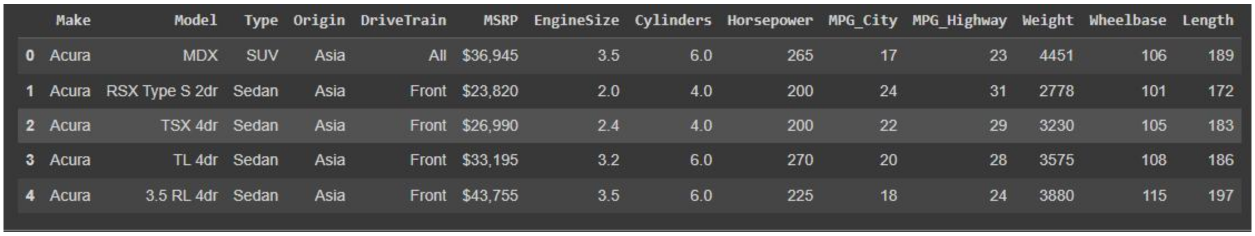
Proposed Methodology

As we have read many articles on car price prediction, we have analyzed and observed that for this type of data set, linear regression was the best suited algorithm. So, we have used Kaggle for data set featuring 13 columns. We cleaned the data set and replaced the columns with the different category values of the column to make them representable with the binary numbers.

We have used the pandas library to read the data from csv file. After dealing with the data set, we moved forward to apply linear regression algorithms implemented in the scikit learn like linear regression, Random Forest Regression, XGBoost Regression.

Dataset Details

There are 13 features used in the model, some of them are listed as follows:

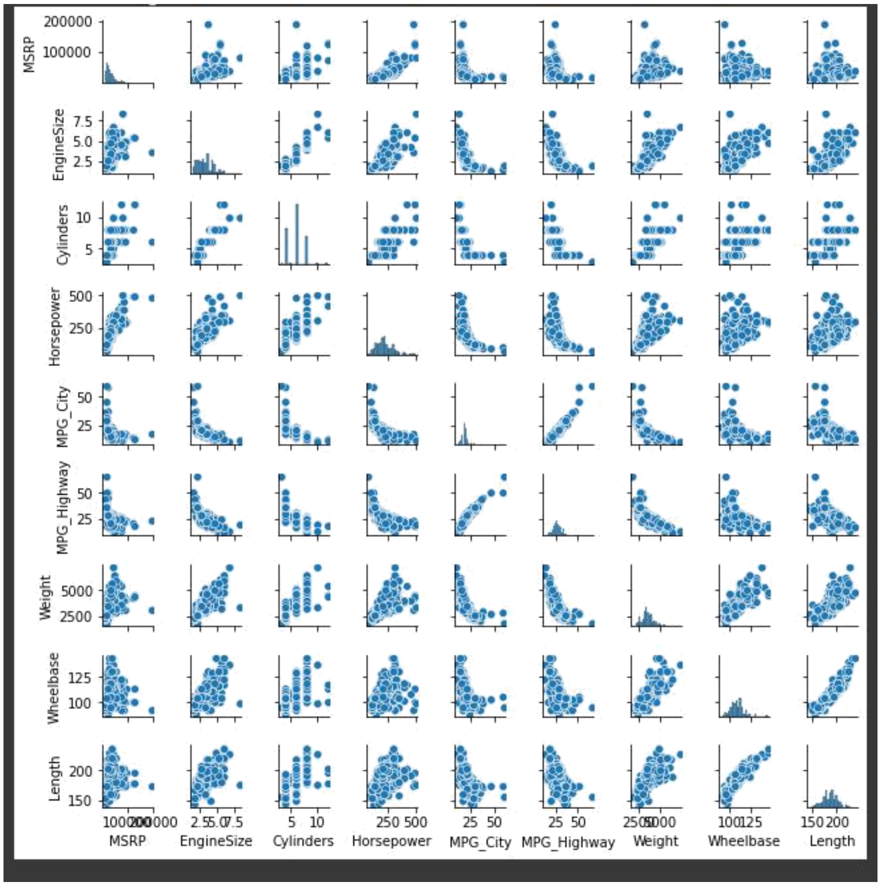


	Make	Model	Type	Origin	DriveTrain	MSRP	EngineSize	Cylinders	Horsepower	MPG_City	MPG_Highway	Weight	Wheelbase	Length
0	Acura	MDX	SUV	Asia	All	$36,945	3.5	6.0	265	17	23	4451	106	189
1	Acura	RSX Type S 2dr	Sedan	Asia	Front	$23,820	2.0	4.0	200	24	31	2778	101	172
2	Acura	TSX 4dr	Sedan	Asia	Front	$26,990	2.4	4.0	200	22	29	3230	105	183
3	Acura	TL 4dr	Sedan	Asia	Front	$33,195	3.2	6.0	270	20	28	3575	108	186
4	Acura	3.5 RL 4dr	Sedan	Asia	Front	$43,755	3.5	6.0	225	18	24	3880	115	197

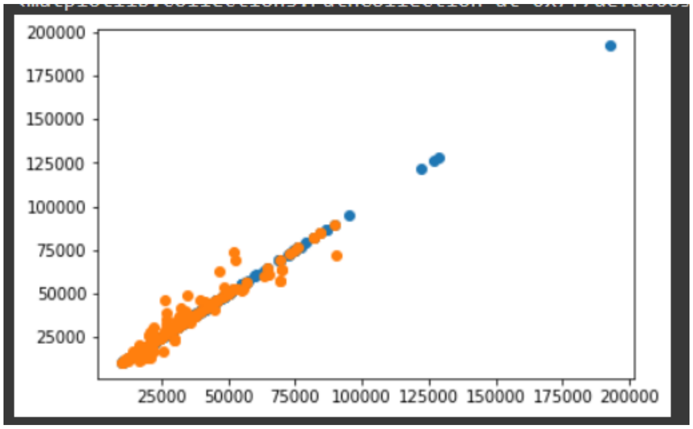
- **Engine Size**
- **Cylinders**
- **Horse Power**
- **MPG City**
- **MPG Highway**
- **Weight**
- **WheelBase**
- **Length**
- **Drive Train**
- **Origin**
- **Type**
- **Model**
- **Make**

Experiments and Results

- Before applying any algorithm, we have observed the correlation of our data as given below.

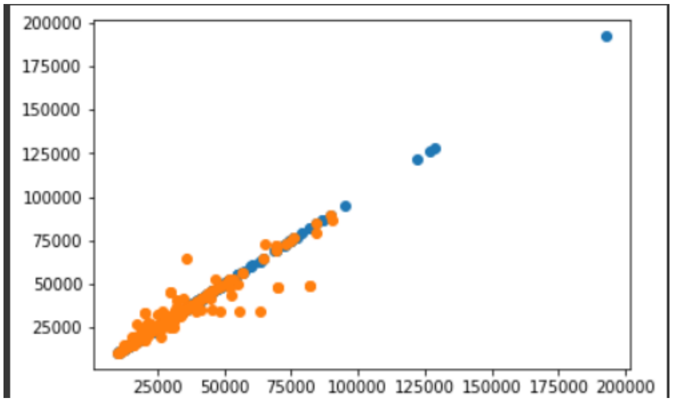


- As it is a linear regression problem, we have applied the linear regression from scikit learn and noticed the score achieved by the model is 93.79%



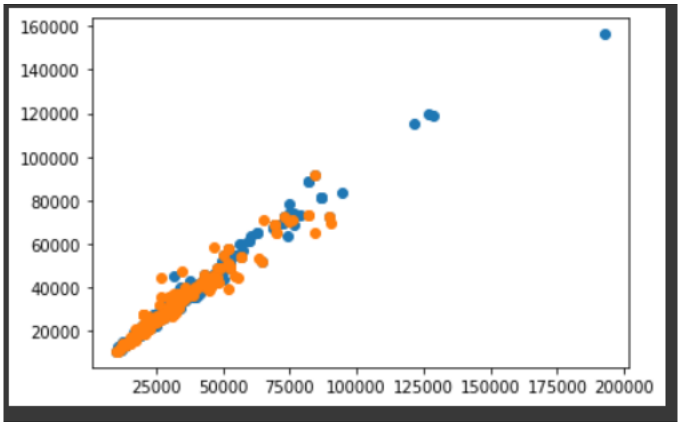
Training and Testing Data by linear Regression

- Since linear regression assumes a linear relationship between the input and output variables, it fails to fit complex datasets properly. In most real-life scenarios, the relationship between the variables of the dataset isn't linear and hence a straight line doesn't fit the data properly.
- We used the Decision Trees here the goal is to create a model that predicts a target variable by using a tree-like pattern of decisions



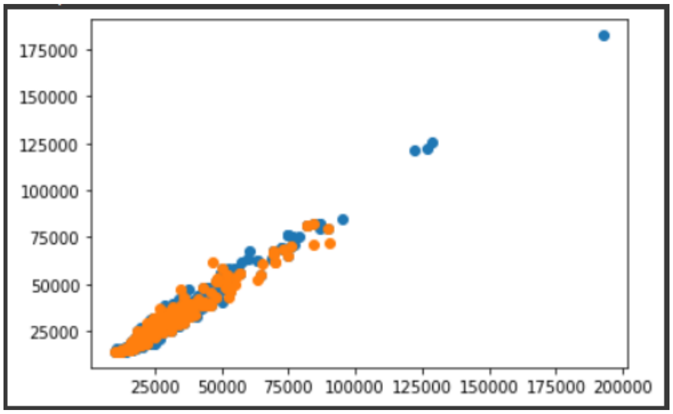
Training and Testing by Decision Trees

- A small change in the data can cause a large change in the structure of the decision tree causing instability. For a Decision tree sometimes, calculation can go far more complex compared to other algorithms. Decision tree often involves higher time to train the model. So, following these imitations of Decision Trees our model predicts with score of 83.3%.
- Random Forest can perform both regression and classification tasks. A random forest produces good predictions that can be understood easily. It can handle large datasets efficiently. The random forest algorithm provides a higher level of accuracy in predicting outcomes over the decision tree algorithm. This algorithm predicts with the score of 93.59%. Many trees can make the algorithm too slow and ineffective for real-time predictions.



Training and Testing Data by Random Forest

- So xgboost will generally fit training data much better than linear regression, but that also means it is prone to overfitting, we have the score prediction of 93.93% by using this algorithm.



Training and Testing Data by XG boost

Training and Testing Data

The data is split into 70% training and 30% testing.

Conclusion

Concludingly, we have reached at the point we have discussed four regression algorithms and the most accurate results were given by the XG boost regression.

Bibliography

[1] Richardson, "Car Price Prediction Using Linear Regression," p. 11, 21 08 2016.

[2] Parudath, *Car Price Prediction,* Lucknow, 2013.

[3] N. a. Jan, *Prediction using linear regression,* Mumby, 2015.