JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Car Price Prediction Using Machine Learning

Abhinandan Singh Dandotiya[1], Dr. Nidhi Dandotiya[2], Dr. Shashikant Gupta[3], Himanshu sahay[4], Ashi Gupta[5], Sonali Rajawat[6]

Asst. Prof. [1] ITM Gwalior, Asst. Prof.[2], Profressor[3], PG Student[4,5,6] ITM University, Gwalior,

Abstract

This work developed a method for forecasting auto prices using supervised machine learning. The research employed multiple linear regressions, a machine learning prediction method that achieved 98% accuracy. We evaluate the accuracy of our findings by comparing the predicted and actual value under a single label. A variety of factors, including make, model, fuel type, body material, location, and optional equipment (such alloy wheels) all contribute to the estimated prices shown in this document.

Keywords: Multiple linear regression, Car price, regression model

1. Introduction

Predicting car price is an interesting and popular problem. This research paper explores the system for car recommendation. Based on the chosen option like fuel type, capacity, and budget this recommender system will recommend the second hand car models to the user. With the help of machine learning techniques and visualization used in Recommender System, The user will get a lot of options so that, he can choose without any confusion. Inthis recommender system we will add all the details of recommended car along with its market price. In this research, we proposed a straightforward, minimal data set, Decision Tree Regression-based model to predict the pricing of second-hand cars. Here, we have gathered and examined a dataset of used cars. Because used automobiles are more affordable and may be sold again after a few years of use for a profit, the majority of people choose to purchase them.

Numerous variables, including gasoline type, colour, model, mileage, transmission, engine, number of seats, etc., affect the price of second-hand cars. The cost of used cars onthe market will continue to fluctuate. As a result, an evaluation model is needed to forecast the price of second-hand cars. We provide a machine learning-based system for estimating used car pricing automated based on the features of the cars. The structure of this paper is as follows: Related research in the area of used-car price prediction is presented in Section II. The research technique for our study is explained in section III. The performance of each machine learning method is examined in Section IV in order to predict the price of used cars. The conclusion of our study is provided in section V, along with a strategy for future work.

2. Literature review

Predicting the value of old cars has been the subject of a large number of researches. Listian, in her master's thesis publication, argued that a regression model based on Support Vector Machines (SVM) could predict the price of a rental automobile with more precision than multivariate regression or simple multiple regression [1]. Richardson presented an alternative approach in his thesis work [2]. The automakers, build stronger vehicles. A question from Richardson.

Using a technique called multiple regression analysis, it was Hybrids' resale value stays higher for longer than that of regular automobiles. Wu et al. [3] utilized a neuro-fuzzy knowledge-based system to conduct research on automobile pricing forecasting. They thought about the make, year, and power plant. Their prediction model and the simple regression model produced the same results. Gonggie [4] proposed an ANN-based model for estimating the value of a secondhand automobile (Artificial Neural Networks). He considered the make, the predicted lifespan, and the annual mileage. Noor and Jan [5] use multiple linear regression to develop a model that predicts future automobile pricing. There are numerous researcher will explore different algorithm technique for this work. Inthis research applying Machine Learning Algorithms to predict the car price accurately.

The objective of this project is to develop an efficient and accurate model that can estimate the price of used cars.

3. Research Methodology

The source of data is Kaggle.com. For each car the following attributes were captured : car_id, symboling, CarName, fueltype, aspiration, carbody, stroke, doornumber, drivewheel, engine location, wheelbase, car length,width,height,weight,fuel system, price in indian rupees, etc.

When the raw data was gathered and saved, the preprocessing phase was carried out. In our case, we simply substituted the most common value for the characteristics that had unexpected ones. Unpriced cars were dumped.

The majority of records are in kilometres, all cars mileage have been reduced to one kmpl in order to prevent mileage conflicts between various vehicle.

To Convert categorical data values into numeric attributes like(Company, Name, Location, fuel, Transmission and owner) we have used a one hot encoding approach.

car_ID	symboling	CarName	fueltype	aspiration	doornumb	carbody	drivewhe	engineloc	wheelbas	carlength	carwidth	carheight	curbweigh	enginetyp	cylindern	enginesiz
1	3	alfa-rome	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130
2	3	alfa-rome	gas	std	two	convertibl	rwd	front	88.6	168.8	64.1	48.8	2548	dohc	four	130
3	1	alfa-rome	gas	std	two	hatchback	rwd	front	94.5	171.2	65.5	52.4	2823	ohcv	six	152
4	2	audi 100 l	gas	std	four	sedan	fwd	front	99.8	176.6	66.2	54.3	2337	ohc	four	109
5	2	audi 100ls	gas	std	four	sedan	4wd	front	99.4	176.6	66.4	54.3	2824	ohc	five	136
6	2	audi fox	gas	std	two	sedan	fwd	front	99.8	177.3	66.3	53.1	2507	ohc	five	136
7	1	audi 100ls	gas	std	four	sedan	fwd	front	105.8	192.7	71.4	55.7	2844	ohc	five	136
8	1	audi 5000	gas	std	four	wagon	fwd	front	105.8	192.7	71.4	55.7	2954	ohc	five	136
9	1	audi 4000	gas	turbo	four	sedan	fwd	front	105.8	192.7	71.4	55.9	3086	ohc	five	131
10	0	audi 5000	gas	turbo	two	hatchback	4wd	front	99.5	178.2	67.9	52	3053	ohc	five	131
11	2	bmw 320i	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108
12	0	bmw 320i	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2395	ohc	four	108
13	0	bmw x1	gas	std	two	sedan	rwd	front	101.2	176.8	64.8	54.3	2710	ohc	six	164
14	0	bmw x3	gas	std	four	sedan	rwd	front	101.2	176.8	64.8	54.3	2765	ohc	six	164
15	1	bmw z4	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3055	ohc	six	164
16	0	bmw x4	gas	std	four	sedan	rwd	front	103.5	189	66.9	55.7	3230	ohc	six	209
17	0	bmw x5	gas	std	two	sedan	rwd	front	103.5	193.8	67.9	53.7	3380	ohc	six	209
18	0	bmw x3	gas	std	four	sedan	rwd	front	110	197	70.9	56.3	3505	ohc	six	209
19	2	chevrolet	gas	std	two	hatchback	fwd	front	88.4	141.1	60.3	53.2	1488	l	three	61
20	1	chevrolet	gas	std	two	hatchback	fwd	front	94.5	155.9	63.6	52	1874	ohc	four	90
21	0	chevrolet	gas	std	four	sedan	fwd	front	94.5	158.8	63.6	52	1909	ohc	four	90
22	1	dodge ran	gas	std	two	hatchback	fwd	front	93.7	157.3	63.8	50.8	1876	ohc	four	90

Table 1. Processed data set sample in CSV Format

Data processing was performed to filter and and remove some extraneous data from Used Cars data set. To more accurately estimate the sales of used automobiles, the model was trained using the processed data and the KNN method.

Application and Technology Applied

Python is utilised as the preferred programming language to implement the machine learning ideasbecause of its many built-in techniques and package libraries.

The following is a list of the well-known libraries and tools we used for this project:

Numpy

NumPy[3] is a general-purpose library for managing arrays. It provides an extremely quick multidimensional array object along with the ability to interact with these arrays. The foundational Python module for scientific computing, to put it simply. NumPy performs well as a multi-dimensional container of general data in addition to its obvious uses in science. NumPy can quickly and easily interact with a variety of databases thanks to its ability to declare any data-types.

Scipy

SciPy, short for Scientific Python, is a Python library specifically designed for scientific and technical computing. Optimization, linear algebra, integration, interpolation, special functions, fast Fourier transform, signal and image processing, and ODE solvers are just some of the usual tasks that SciPy offers modules for SymPy, pandas, and Matplotlib are all components of the NumPy stack, along with SciPy, which is based on the NumPy array object.

Scikit-learn

Through a standardized Python interface, Scikit-learn offer a variety of supervised and unsupervised learning techniques. It is released under several Linux distributions and is available under a liberal simplified BSD license, which promotes both academic and commercial use.

Jupyter notebook

You can create and share documents with real-time code, equations, graphics, and text using the open-source Jupyter Notebook web tool. It includes data translation, data cleaning, statistical modelling, data visualization, and many other things.

4. Results and discussion

We examine the methodologies and approaches utilized in our module; This incorporates analyses of our dataset using statistical methods using a number of dispersed graphs, a violin graph, comparison charts, and bar graphs to determine the optimal methodology. We first preprocess and cleanse the data in our dataset. After locating them, 15% of the tuples with null values were eliminated.

80 % of the dataset was used for training, and 20 %for testing. We use Decision tree regression for estimating a vehicle's value using the Python Scikit-Learn package. The yearof registration had a marginally greater influence.

Fig.1: Histogram of all attributes

Fig 1.represents the histogram it provides a visual interpretation of numerical data by showing the number of data points that fall within a specified range of values.

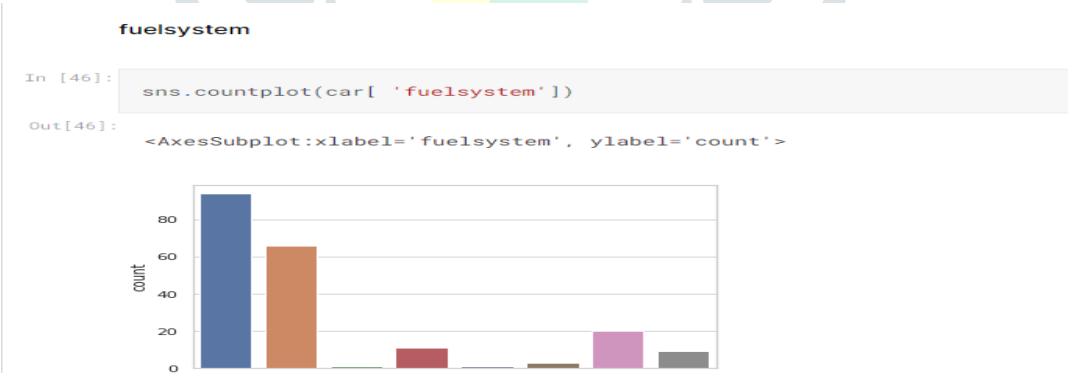

Fig.2: Fuel System of a car

Fig2. Represents the countplot to Show the counts of observations of fuel system in each categorical bin using bars.

```
df2 = pd.get_dummies(df1, columns=categorical, drop_first=True)
df2.head()
```

	symboling	wheelbase	carlength	carwidth	carheight	curbweight	enginesize	boreratio	stroke	compressionra
0	3	88.6	168.8	64.1	48.8	2548	130	3.47	2.68	9.0
1	3	88.6	168.8	64.1	48.8	2548	130	3.47	2.68	9.0
2	1	94.5	171.2	65.5	52.4	2823	152	2.68	3.47	9.0
3	2	99.8	176.6	66.2	54.3	2337	109	3.19	3.40	10.0
4	2	99.4	176.6	66.4	54.3	2824	136	3.19	3.40	8.0

Fig.3: Manipulating the data

Convert a categorical variable into dummy variables.

```
In [9]:
gs_dt.fit(X_train,y_train)

# Store Best Estimator
best_dt_estimates = gs_dt.best_estimator_
best_dt_estimates

Out[9]:
DecisionTreeRegressor(max_depth=6, random_state=42)
```

Fig.4. fit grid search to get best estimator

```
In [10]:
best_dt_estimates.fit(X_train, y_train)

Out[10]:
DecisionTreeRegressor(max_depth=6, random_state=42)
```

Fig.5. fit model with best estimator

```
In [11]:
y_train_predicted = best_dt_estimates.predict(X_train)
y_test_predicted = best_dt_estimates.predict(X_test)
```

Fig.6 Make prediction

Check Performance

In [12]:
```
performance_dict = {
        'Model_Name' : gs_dt.best_estimator_,
        'Train_RMSE' : round(mean_squared_error(y_train, y_train_predicted, squared=Fals
e),2),
        'Test_RMSE'  : round(mean_squared_error(y_test, y_test_predicted, squared=False),
2)
    }
```

In [13]:
```
performance = pd.DataFrame([performance_dict])
performance
```

Out[13]:

	Model_Name	Train_RMSE	Test_RMSE
0	DecisionTreeRegressor(max_depth=6, random_stat...	988.61	2462.68

Fig.7 Check train and test performance

Conclusion

We go over the findings and observations we made while putting this module into practice in this chapter. We used well-known Python libraries' algorithms to construct machine learning algorithmic paradigms successfully. On our dataset, we first pre-process it and clean up the data. 15% of the tuples were determined to have null values, thus we trimmed those tuples. The findings revealed a positive link between price and miles travelled a negative link between price and year of registration, but no association between mileage and registration year. While negative correlation pertains to the idea of inverse proportion, positive correlation is essentially related to the idea of direct proportion. The model was trained using three lakh tuples. The year of registration had a marginally greater influence. On two separate vehicle models, Decision Tree Regression and Classification and Regression Trees (CART) arecontrasted.

Future work will focus on the variable options over the project's algorithms that were employed. Only could we compare two algorithms while several other algorithms are explored. Things already exist and could be more precise. A system or service will add more spaces. More price accuracy in the system, i.e.

1) Horsepower
2) Battery power
3) Suspension
4) Cylinder
5) Torque

As we are all well aware, technology is constantly evolving, and the technology of cars has also advanced, so our next hybrid vehicles, electric vehicles, and driverless vehicles will beupgraded cars.

Refrences

1.M. Antonakakis, Understanding the mirai botnet in *Proc. of USENIX Security Symposium*, 2017.

2.Shiravi, Toward developing a systematic approach to generate benchmark datasets for intrusion detection *Comput. Sec* 31 3 357374

3.Gegic Car price prediction using machine learning techniques *TEM J* 8, no. 1

4.Information regarding machine learning techniques and algorithmshttps: en.wikipedia.org/wiki/Machine_learning.

5.R. Ragupathy and L. Phaneendra Maguluri, "Comparative analysis of machine learning algorithms on social media test," *Int J Eng Technol(UAE)*, vol. 7

6.

7.R. Doshi, Machine earning DDoS etection for onsumer Internet of hings evices, 2018 IEEE Security and Privacy Workshops (SPW) San Francisco, CA, 2018, pp. 29-35

8.Chen, Yi-Wen, Jang-Ping Sheu, Yung-Ching Kuo, and Nguyen Van Cuong. "Design and implementation of IoT DDoS attacks detection system based on machine learning." In *2020 European Conference on Networks and Communications (EuCNC)*, pp. 122-127. IEEE, 2020.

9.Ashraf, A. and Elmedany, W.M., 2021, October. IoT DDoS attacks detection using machine learning techniques: A Review. In *2021 International Conference on Data Analytics for Business and Industry (ICDABI)* (pp. 178-185). IEEE.

10. Malik, Manisha, and Maitreyee Dutta. "Feature Engineering and Machine Learning Framework for DDoS Attack Detection in the Standardized Internet of Things." *IEEE Internet of Things Journal* (2023).