International journal of basic and applied research
www.pragatipublication.com
ISSN 2249-3352 (P) 2278-0505 (E)
Cosmos Impact Factor-5.86

PREDICTION THE PRICE OF USED CAR USING MACHINE LEARNING

[1]Dr. N. Baskar,[2]V. Srija,[3]S. Aakanksha,[4]U. Sruthi
[1]Professor,[234]Students
Department Of CSE
Malla Reddy Engineering College for Women

ABSTRACT:

Due to the obvious level of expertise and effort required, the prediction of automobile prices has been a popular topic of study. In order to make a solid and precise forecast, a large number of different characteristics are researched. In order to construct a model that can foretell the value of pre-owned automobiles. Machine learning techniques such as logistic regression, Random Forest, Support Vector Machine, and linear regression were used. But the aforementioned methods were used in conjunction with one another to do regressions and ensemble work. The information that went into making the forecast was gathered Then, in order to determine which method is most suited to the given data set, their respective results were compared. Predicting the optimal percentage and revealing the outcome with the appropriate algorithms are both tasks that need the use of algorithms.

I. INTRODUCTION

1.1 OVER VIEW

The challenge of estimating the value of pre-owned automobiles is intriguing and crucial. The National Transport Authority reports that there was a 234% surge in the number of vehicles registered between 2003 and 2013. There were 68,524 vehicles registered in 2003; now, there are 160,701. Used automobiles and reconditioned foreign vehicles are anticipated to see a surge in sales as a result of the current economic climate. Sales of new automobiles fell by 8 percent in 2013, according to [2].

Leasing an automobile instead of owning one altogether is popular in many industrialized nations. The parties to a lease agree that the buyer will pay the seller or third party (often an insurance company, bank, or other financial institution) a certain amount of money over a certain period of time (often a set number of months or years). The buyer gets the option to purchase the vehicle at its residual value, which is its anticipated resale value, after the lease term ends. Therefore, being able to accurately forecast the salvage value (residual value) of vehicles is of economic significance to seller/financers. If the seller or financier first underestimates the residual value, the customers will have to pay more in installments and will likely go elsewhere for their financing needs. Overestimating the residual value will result in reduced installments for customers, but it will make it difficult for the vendor or financier to sell the expensive used automobiles at that price. Therefore, it is clear that determining the value of pre-owned vehicles is also of great significance from a business perspective. Due to an error in estimating the residual value of leased vehicles, German manufacturers lost one billion euros in the American market [3]. When purchasing a new automobile in Mauritius, the majority of buyers are understandably wary about the vehicle's potential resale value on the used-car market after a certain amount of time has passed.

The art of estimating a car's future worth is complex. Used automobile values are dependent on a variety of variables, which is common information. Information on the vehicle's age, make and model, country of origin, mileage, horsepower, and total kilometers driven are often the most relevant.

gasoline efficiency is especially crucial in light of the recent spike in gasoline costs. Sadly, in reality, the vast majority of drivers are clueless about their vehicle's fuel consumption per kilometer. The price can be affected by a lot of other things, like the fuel type, the interior style, the braking system, acceleration, the engine size (in cc), the safety index, the car's size, the number of doors, the color of the paint, the car's weight, customer reviews, the manufacturer's prestigious awards, the car's physical condition, the presence or absence of cruise control, the type of transmission (automatic or manual), the owner's status (individual or corporate), and other options like air conditioning, sound system, power steering, cosmic wheels, and GPS navigator. Buyers in Mauritius place a premium on knowing the car's history, whether it has been in any major accidents, and if it was driven by a woman. The car's aesthetics are a major factor in the final pricing. Numerous variables affect the pricing, as we can see. Regrettably, it is not always possible to get information on all of these issues, so the buyer is left with limited options when deciding on a price.

II. LITERATURE SURVEY

Predicting the value of old automobiles has been the subject of several research. It is common practice for researchers to use historical data to predict product pricing. In order to forecast the pricing of used automobiles in Mauritius, Pudaruth utilized a combination of multiple linear regression, k-nearest neighbours, Naive Bayes, and decision tree approaches. The cars in question were not brand new. It was shown that the pricing from different tactics are quite comparable when the forecast results are compared. Both the decision-tree method and the Naive Bayes approach failed miserably when it came to numerical value classification and prediction.

The study conducted by Pudaruth indicates that the limited sample size does not provide accurate predictions.

[2] in For the purpose of numerical value classification and prediction, Kuiper, S. (2008) presented a multivariate regression model. It shows how to use this multivariate regression model to predict the GM car prices for 2005. It is not necessary to have any particular information in order to forecast automobile prices. There is sufficient information to make price predictions from the web. In order to determine which factors were most important to include in the model, the article's author used variable selection strategies while making the same automobile price estimate.

As a strategy for estimating the pricing of secondhand automobiles using Random Forest, Pal et al. [3] found in 2019. Using the Kaggle dataset, which yielded an accuracy of 95% for train-data and 83.62% for test-data, the article assessed the accuracy of used-car pricing prediction. Price, kilometer, brand, and vehicle type were determined to be the most important characteristics for this forecast after removing outliers and irrelevant data. Random Forest, being an advanced model, outperformed previous efforts utilizing these datasets in terms of accuracy.

Table 4 The need to develop a model for predicting the value of pre-owned vehicles in Bosnia and Herzegovina is highlighted by Gegic, E. et al. (2019). A number of machine learning algorithms were used, including ANNs, SVMs, and RFs. Nevertheless, a combination of the aforementioned strategies was used. The data for the prediction was retrieved from the website autopijaca.ba using a web scraper that was developed in the PHP computer language. Then, several algorithms' performances were evaluated to find the one that worked best with the given data. The finished prediction model was stored in a Java application. The model was also shown to have an accuracy of 87.38% when

tested with real data. In 2019, a machine learning-based approach to auto-resales was shown by Dholiya et al.

[5] Dholiya, M., et al. designed their system with the user's budget in mind, so they can provide an accurate estimate of the vehicle's potential cost. This web-based system may also provide the user with a list of possible vehicle types based on the details of the vehicle they are searching for. In doing so, it helps to provide the vendor or buyer with relevant data upon which to build their choice. The multiple linear regression approach is used to generate predictions by this system. The model was trained using data collected over a long period of time. At first, the KDD (Knowledge Discovery in Databases) procedure was used to collect the raw data. Previous to that, material was cleaned and preprocessed to find useful patterns, and from there, some significance was extracted.

III. METHODOLOGY

Using Kaggle, I was able to choose the necessary dataset for used vehicle pricing together with the characteristics and parameters. Data scientists and analysts may access data and notebooks on the open-source Kaggle platform, which specializes in machine learning. Prior to implementing any algorithm for price prediction, the necessary data is cleaned and pre-processed using machine learning methods. Following data cleaning and pre-processing, we divide the data into two sets, one for training and the other for validation, using the train and test sets, respectively. After that, we need to test and train a basic linear regression model using the roc auc score to forecast the output, and then we need to train and test it again using a multiple linear regression model to evaluate its accuracies. Next, in order to forecast the final vehicle price, we must use clustering, logistic regression, and k-nearest neighbours algorithms. We can also use decision-tree and random-forest methods. Before making a final decision, we must evaluate each machine learning algorithm's accuracy and choose the most suitable one for the forecast.

IV. MODELING AND ANALYSIS

We conduct a little linear regression in between these cases.

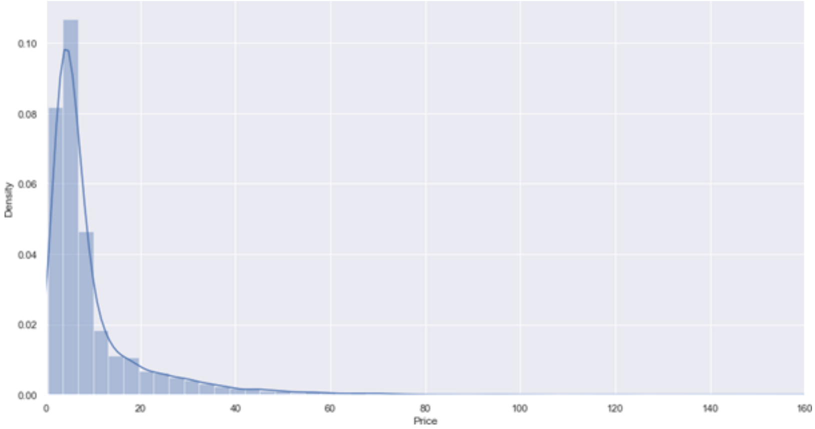


Figure 1: Comparisons of price and density

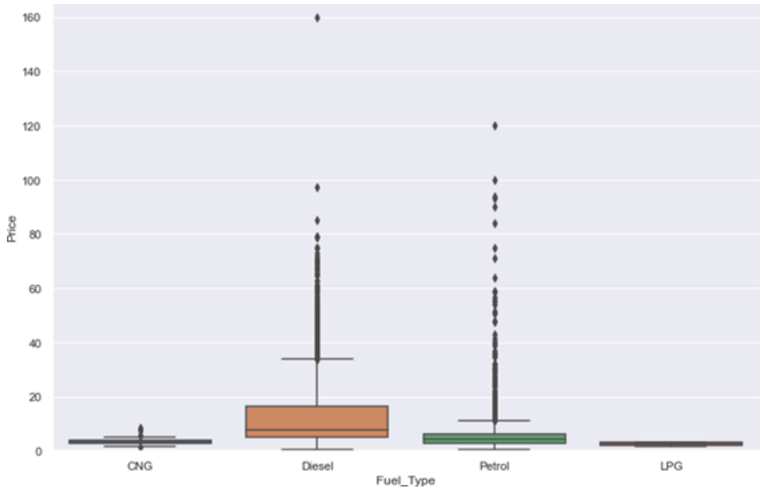


Figure 2: Box plot between various fuel types of cars.

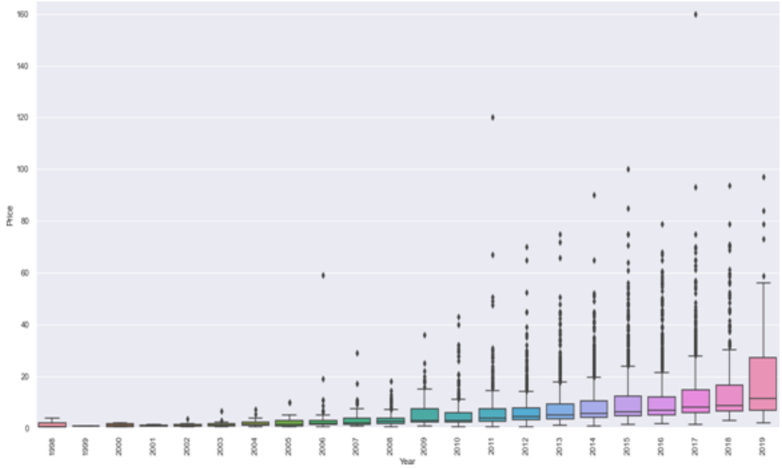


Figure 3: Box plot between various years of cars.

V. RESULTS AND DISCUSSION

The pair plot makes it clear that the following characteristics have the most impact: price, engine type, power, mileage, fuel type, and gearbox type. Consequently, we must construct the model using those attributes. A primary objective of this research is to provide a method for both online and offline automobile dealers and buyers to anticipate future car prices. Kilometers, gearbox, owner type, mileage, engine cc, and other characteristics are used to forecast used vehicle values. In the future, this vehicle can be fine-tuned using hyperparameter tuning and the use of neural networks and other deep learning methods, so the costs shown here are accurate and trustworthy.

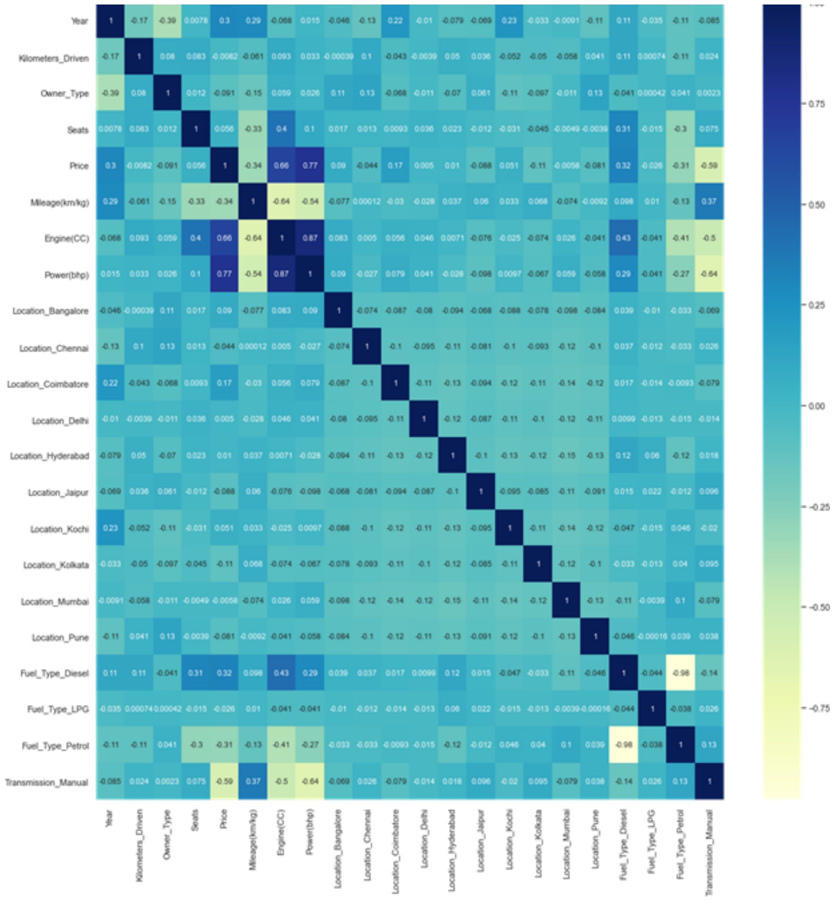
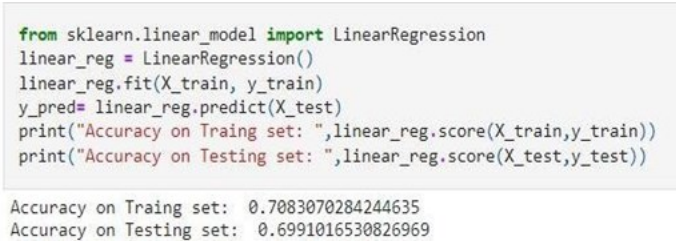


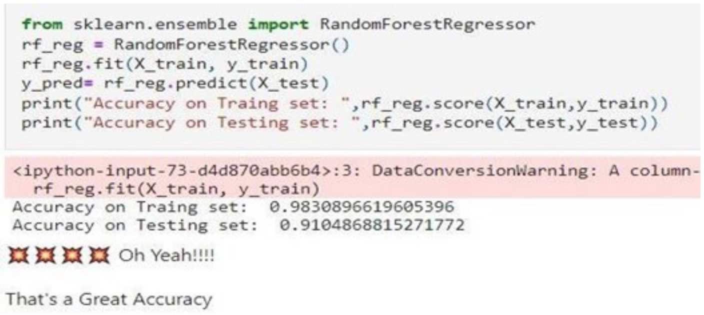
Figure 4: Pair plot between all the features.



```
from sklearn.linear_model import LinearRegression
linear_reg = LinearRegression()
linear_reg.fit(X_train, y_train)
y_pred= linear_reg.predict(X_test)
print("Accuracy on Traing set: ",linear_reg.score(X_train,y_train))
print("Accuracy on Testing set: ",linear_reg.score(X_test,y_test))
```

```
Accuracy on Traing set:  0.7083070284244635
Accuracy on Testing set:  0.6991016530826969
```

Figure 5: Accuracy of linear regression model



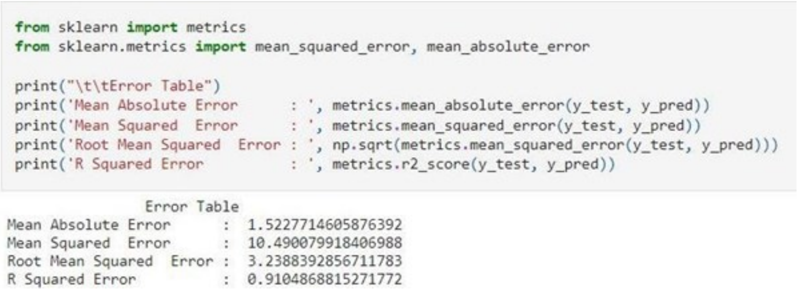
```
from sklearn.ensemble import RandomForestRegressor
rf_reg = RandomForestRegressor()
rf_reg.fit(X_train, y_train)
y_pred= rf_reg.predict(X_test)
print("Accuracy on Traing set: ",rf_reg.score(X_train,y_train))
print("Accuracy on Testing set: ",rf_reg.score(X_test,y_test))
```

```
<ipython-input-73-d4d870abb6b4>:3: DataConversionWarning: A column-
  rf_reg.fit(X_train, y_train)
Accuracy on Traing set:  0.9830896619605396
Accuracy on Testing set:  0.9104868815271772
```

💥💥💥💥 Oh Yeah!!!!

That's a Great Accuracy

Figure 6: Accuracy of random forest



```
from sklearn import metrics
from sklearn.metrics import mean_squared_error, mean_absolute_error

print("\t\tError Table")
print('Mean Absolute Error      : ', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared  Error      : ', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared  Error : ', np.sqrt(metrics.mean_squared_error(y_test, y_pred)))
print('R Squared Error          : ', metrics.r2_score(y_test, y_pred))
```

```
                Error Table
Mean Absolute Error      :  1.5227714605876392
Mean Squared  Error      :  10.490079918406988
Root Mean Squared  Error :  3.2388392856711783
R Squared Error          :  0.9104868815271772
```

Figure 7: Accuracy and errors of random forest model

By comparing the linear regression model's accuracy score to that of the random forest model, we find that the latter performs much better, with an accuracy score of 91%. Hence, a random forest model is the way to go for a project involving the prediction of automobile prices.

VI. CONCULSUION

The large number of factors that need to be examined in order to make an accurate projection of a car's pricing makes it a difficult undertaking. Gathering and cleaning up the data is the first and most important stage in making predictions. Clean and normalize data in this study so that machine learning algorithms don't have to deal with extraneous noise. Although data cleaning is an effective method for improving prediction performance, it is not enough to handle complicated data sets like the one used in this study. The accuracy dropped below 50% when just one machine algorithm was applied on the dataset. Hence, a greater number of algorithms is required for more accurate percentage prediction. When contrasted with the strategy of using only one machine learning technique, this is a huge increase. One problem with the suggested approach is how much more computing power it needs than just one machine learning method. Despite the system's impressive performance in the automobile price prediction task, our goal for future study is to see whether it can effectively handle other types of data.

REFERENCES

[1] NATIONAL TRANSPORT AUTHORITY. 2014. Available from: http://nta.gov.mu/English/Statistics/Pages/Archive.aspx [Accessed 15 January 2014]

[2] MOTORS MEGA. 2014. Available from: http://motors.mega.mu/news/2013/12/17/auto-market8- decrease-sales-newcars/ [Accessed 17 January 2014].

[3] LISTIANI, M., 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Thesis (MSc). Hamburg University of Technology.

[4] Oprea, C, Making the decision on buying second-hand car market using data mining techniques(Special, 2010), pp.17-26.

[5] C Ozgur, Z Hughes, G Rogers and S Parveen, Multiple Linear Regression Applications Automobile Pricing (International Journal of Mathematics and Statistics Invention, 2016), pp.01-10

[6] Lessmann, Stefan, M. Listiani, and S. Voß, Decision support in car leasing: A forecasting model forresidual value estimation (2010).

[7] G.Chandrashekar and F. Sahin, “A survey on feature selection methods,” Computers Electrical Engineering, vol. 40, no. 1, pp. 16–28, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0045790613003066

[8] M.C.Newman,“Regression analysis of log-transformed data: Statistical bias and its correction,” Environmental Toxicology and Chemistry, vol. 12, no. 6, pp. 1129–1133, 1993. [Online]. Available: http://dx.doi.org/10.1002/etc.5620120618

[9] R.Taylor, “Interpretation of the Correlation Coefficient: A Basic Review,” Journal of Diagnostic Medical Sonography, vol. 6, no. 1, pp. 35–39, 1990

[10] Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques”;(IJICT 2014)

[11] Enis gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric, ”Car Price Prediction Using Machine Learning”; (TEM Journal 2019)

[12] Ning sun, Hongxi Bai, Yuxia Geng, Huizhu Shi, “Price Evaluation Model In Second Hand Car System Based On BP Neural Network Theory”; (Hohai University Changzhou, China)

[13] Nitis Monburinon, Suwat Rungpheung, Sabir Buya, Pitchayakit Boonpou, “Prediction of Prices for Used Car by using Regression Models” (ICBIR 2018).