Second-hand Car Price Prediction Using Machine Learning

Albin P Thomas
Amal Jyothi College of Engineering
APJ Abdul Kalam University
Kottayam, Kerala,India
albinpthomas@mca.ajce.in

Shelly Shiju George
Assistant Professsor, ,Department of Computer Application
Amal jyothi College Of Engineering,
Kanjirappally,Kerala
shellyshijugeorge@amaljyothi.ac.in

***Abstract*— A new car's price is established by the manufacturer, with the government paying some additional costs in the form of taxes. Customers who buy a new car may feel assured that their money will be wisely spent. However, due to rising new car prices and purchasers' inability to afford them, global used car sales are expanding. As a result, a Used Car Price Prediction system that can reliably appraise a car's worth based on a variety of characteristics is in high demand. We look at how supervised machine learning techniques may be used to predict the price of used cars in this study. The prediction is based on historical data gleaned from daily newspapers. Several methods were used to construct the predictions, including multiple linear regression analysis, Random Forest Regression, Random Forest Regression, and Randomized Search cv. We will use several characteristics such as Present Price, Selling Price, kilometers Driven, Fuel Type, Year, and others to anticipate the price of used cars in this Project. Kaggle was utilized to obtain the data for this Project. After then, the forecasts are analyzed and compared to see which ones provide the best outcomes. A seemingly easy problem turned out to be rather difficult to solve precisely. All four techniques had similar outcomes. So the model helps to predict cars actual price rather than predicting car's price range. A user interface has also been created that collects input from any user and displays the price of a car based on the information provided by the user.**

****Knowing how to evaluate the market worth of used cars may be beneficial to both buyers and sellers****

There are lots of individuals who are interested in the used car market at some points in their life because they wanted to sell their car or buy a used car. In this process, it's a big corner to pay too much or sell less then its market value. In the future, we intend to use more complicated algorithms to generate predictions.

Keywords: Jupyter, Linear regression analysis, Random Forest Regression, Random Forest Regression, and Randomized Search CV

I. INTRODUCTION

Predicting the price of old automobiles is a difficult but fascinating topic.[1] The forecasts are based on data gathered from daily publications throughout time. The predictions were made using a variety of approaches, including multiple linear regression analysis, k-nearest neighbors, nave bayes, and decision trees. Here we use linear regression and random forest.[1] According to data gathered from the National Transportation Authority, the number of automobiles registered increased by 234 percent between 2003 and 2013. The number of automobiles registered has increased from 68, 524 in 2003 to 160, 701. With the current economic situation, it is anticipated that sales of old automobiles and second-hand imported (reconditioned) cars would grow. According to reports, the number of new automobile sales fell by 8% in 2013. It is usual in many industrialized nations to lease an automobile rather than buy one entirely. A lease is a legally binding agreement between a buyer and a seller (or a third party – commonly a bank, insurance business, or other financial institution) in which the buyer agrees to pay the seller/financer regular monthly or annual payments. After the lease time has ended, the buyer has the option to purchase the vehicle at its residual value, or estimated resale value. As a result, being able to accurately anticipate the salvage value (residual value) of automobiles is of commercial significance to sellers and financiers. The payments will be higher if the seller/financer first underestimates the residual value Customers will very certainly seek for another seller/financer as a result.[5] The instalments will be cheaper for the customers if the residual value is over-estimated, but the seller/financer may have a tough time selling these high-priced used automobiles at the over-estimated residual value. As a result, we can see that calculating the price of used automobiles is also highly important from a business standpoint.[1] Because of a miscalculation of the residual value of leased automobiles, German automakers lost one billion euros in the United States.[3] By using statistical models to anticipate pricing, it is possible to obtain an approximate estimate of the price without having to enter the specifics into the desired website.

II. RELATED WORKS

The study on estimating the price of secondhand automobiles is new, but it is also scarce. In MSc thesis that a regression mode based on support vector machines (SVM) may more accurately predict the residual price of leased automobiles than simple multiple regression or multivariate regression. SVM is better at dealing with data with a lot of dimensions and avoids both overfitting and underfitting. The study's sole flaw is that the advantage of SVM regression over basic regression was not measured in simple terms like mean deviation or variance.

A novel artificial neural network-based methodology for predicting the residual value of privately owned secondhand automobiles The study's major elements were mileage, manufacturer, and estimated useful life. The model was tweaked to accommodate nonlinear connections, which are difficult to analyse using traditional linear regression

DOI: 10.5281/zenodo.6364937
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Proceedings of the National Conference on Emerging Computer Applications (NCECA)-2022
Vol.4, Issue.1

approaches. This model was proven to be fairly reliable in predicting the residual value of old autos.

III. METHOD OF PREDICTION

The required data is collected from the online second-hand markets, newspapers and showrooms. We collected all the data in the interval of 2003-2018 which made a huge impact on car price. The system comprises two basic phases: 1. Training phase: The system is trained by using the data in the data set and fitting a model (line/curve) depending on the algorithm chosen. 2. Testing phase: the system is fed inputs and checked to ensure that it works properly. The precision is verified. As a result, the data used to train or test the model must be suitable. Because the system is intended to identify and anticipate the price of a used automobile, proper algorithms must be employed to do the two distinct jobs. Different algorithms were examined for accuracy before they were chosen for future usage. The best one for the job was picked. Seasonal patterns don't affect the price of car too much. We constructed a histogram to help us understand the data better. Due to the high price sensitivity of used cars, we noticed that the dataset had a significant number of outliers. Vehicles with little mileage from the previous year often sell for a higher price; however, several data points contradicted this. This is due to the fact that the car's accident history and present condition may have a significant influence on its cost. Because we can't have access to vehicle history or condition, we limited our data to three standard deviations around the mean to eliminate outliers.

The following key features of the car is collected for analysis:

- Car name
- Year
- On-road price
- Total kilometers completed
- Method of Transmission

Figure 1: Sample Data

Car_Name	Year	Selling_Pri	Present_P	Kms_Drive	Fuel_Type	Seller_Typ	Transmiss	Owner
ritz	2014	3.35	5.59	27000	Petrol	Dealer	Manual	0
sx4	2013	4.75	9.54	43000	Diesel	Dealer	Manual	0
ciaz	2017	7.25	9.85	6900	Petrol	Dealer	Manual	0
wagon r	2011	2.85	4.15	5200	Petrol	Dealer	Manual	0
swift	2014	4.6	6.87	42450	Diesel	Dealer	Manual	0
vitara brez	2018	9.25	9.83	2071	Diesel	Dealer	Manual	0
ciaz	2015	6.75	8.12	18796	Petrol	Dealer	Manual	0
s cross	2015	6.5	8.61	33429	Diesel	Dealer	Manual	0
ciaz	2016	8.75	8.89	20273	Diesel	Dealer	Manual	0

A. Techniques Used

- Linear Regression Algorithm

Because of its simplicity and relatively short training period, Linear Regression was chosen as the initial model. The feature vectors were generated straight from the features, without any feature mapping. Since the results were clearly low variance, no regularization was applied. Linear regression algorithm is used for supervised learning in Machine learning.It predicts a value based on the given dataset. It represents relationship between variables and output.

```
from sklearn.linear_model import LinearRegression
lr = LinearRegression()

lr.fit(x_train,y_train)

y_pred_lr = lr.predict(x_test)

r_squared = r2_score(y_test,y_pred_lr)
rmse = np.sqrt(mean_squared_error(y_test,y_pred_lr))
print("R_squared :",r_squared)
```

- Random Forest Regression

It is an ensemble learning method. Ensemble learning is used for combining results to get more accurate result. Ensemble learning underpins the Random Forest regression model. It builds the ensemble model using a decision tree model, which, as the name indicates, is made up of multiple decision trees. The benefit of this strategy is that the trees are created in parallel and are extremely uncorrelated, resulting in good results because each tree is not sensitive to specific errors from other trees. This uncorrelated behavior is helped by Bootstrap Aggregation or bagging, which provides the randomness required to form robust and uncorrelated trees. This model was created to account for the large number of features in the dataset and to compare a bagging technique to the gradient boosting methods that come after.

```
from sklearn.ensemble import RandomForestRegressor
rf = RandomForestRegressor()

# Training Model
rf.fit(x_train,y_train)

# Model Summary
y_pred_rf = rf.predict(x_test)

r_squared = r2_score(y_test,y_pred_rf)
rmse = np.sqrt(mean_squared_error(y_test,y_pred_rf))
print("R_squared :",r_squared)
```

DOI: 10.5281/zenodo.6364937
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Proceedings of the National Conference on Emerging Computer Applications (NCECA)-2022
Vol.4, Issue.1

- Gradient Boosting Regressor

  Gradient Boosting is a decision tree-based approach characterized as "a means of transforming poor learners into strong learners." This means that, similar to a standard boosting strategy, observations are assigned various weights, and the weights of difficult-to-predict observations are raised based on certain metrics, and the weights of difficult-to-predict observations are sent into another tree to be trained. The gradient of the loss function serves as the measure in this case. This model was used to accommodate for non-linear correlations between attributes and estimated price by dividing the data into 100 regions.
  The difference between the current forecast and the known accurate target value is calculated. The term "residual" refers to this disparity.

```
from sklearn.ensemble import GradientBoostingRegressor
gbt = GradientBoostingRegressor()


gbt.fit(x_train,y_train)


y_pred_gbt = gbt.predict(x_test)

r_squared = r2_score(y_test,y_pred_gbt)
rmse = np.sqrt(mean_squared_error(y_test,y_pred_gbt))
print("R_squared :",r_squared)
```

- RandomizedSearch CV

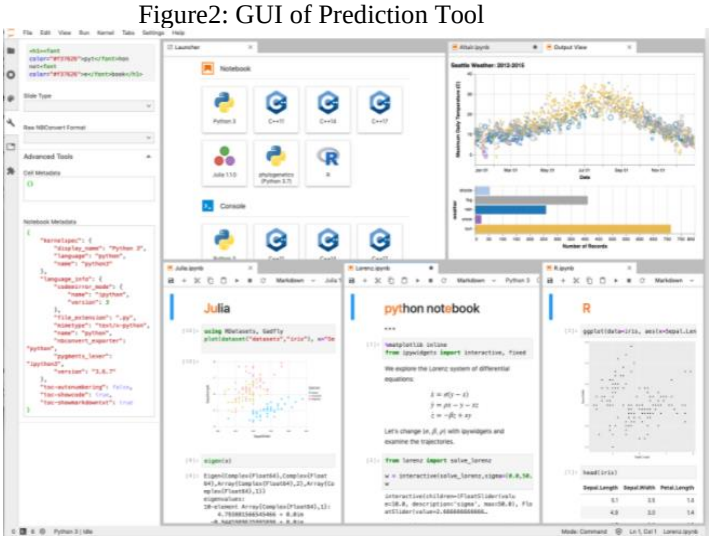
  RandomizedSearchCV has two methods: "fit" and "score." Cross-validated search over parameter settings is used to improve the parameters of the estimator used to apply these approaches.

```
from sklearn.model_selection import RandomizedSearchCV

n_estimators = [int(x) for x in np.linspace(start = 100, stop = 1200, num = 12)]

max_features = ['auto', 'sqrt']

max_depth = [int(x) for x in np.linspace(5, 30, num = 6)]
e
min_samples_split = [2, 5, 10, 15, 100]

min_samples_leaf = [1, 2, 5, 10]

random_grid = {'n_estimators': n_estimators,
               'max_features': max_features,
               'max_depth': max_depth,
               'min_samples_split': min_samples_split,
               'min_samples_leaf': min_samples_leaf}

print(random_grid)
```

- Jupyter

  Jupyter is a computational notebook, which is a free, open-source, interactive web application that allows academics to mix software code, computational output, explanatory text, and multimedia resources in a single document. Users may design and arrange workflows in data science, scientific computing, computational journalism, and machine learning using the platform's versatile interface. Extensions that increase and enhance functionality are possible with a modular architecture.

Figure2: GUI of Prediction Tool



IV. RESULTS OF THE PREDICTION

After we put the data into the various algorithms to the output is predicted using the variables like car name, model year, total kilometers and mode of transmission comparing with the given data set. The algorithm calculates a sum value for the car by assigning different scores to each variable. The method iterates over the variables to get the optimal result that is beneficial to both the seller and the buyer.

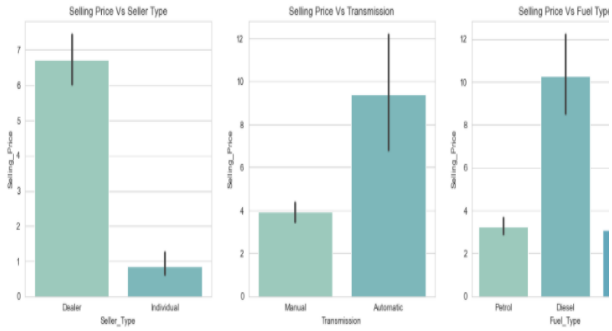
The result we got is:

People pay close attention to the odometer value of a secondhand automobile before purchasing it. We can see that the odometer has a major impact on the pricing of an automobile. However, this does not imply that only automobiles with little mileage are available for purchase. High-odometer automobiles can be purchased depending on the pricing (Figure 4). In addition, the most common used automobiles have an odometer reading of roughly 100,000 miles. Many automobiles are available till the odometer reaches 150,000.

DOI: 10.5281/zenodo.6364937
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

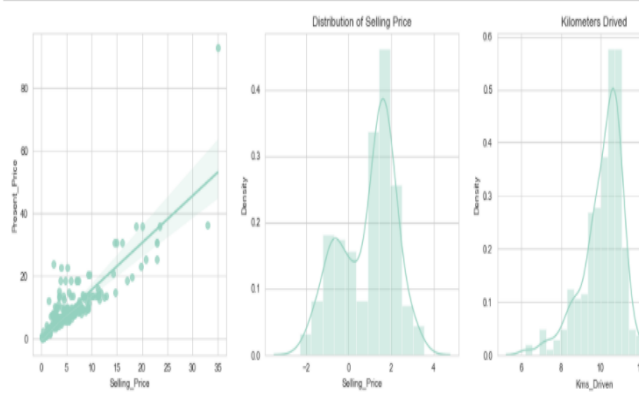
Proceedings of the National Conference on Emerging Computer Applications (NCECA)-2022
Vol.4, Issue.1

Figure3: Bar Plot



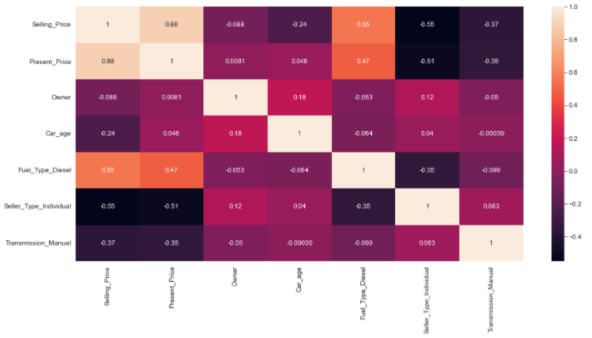
On the secondhand automobile market, the car's manufacturer is another essential factor. One of the most well-known manufacturers is Maruti Suzuki. As major automakers, Toyota and Hyundai are next in line. It is clear that Japanese automobiles account for a significant portion of the used car market. American automobiles, on the other hand, continue to be in high demand and have a commanding position.

Figure4: Dist Plot

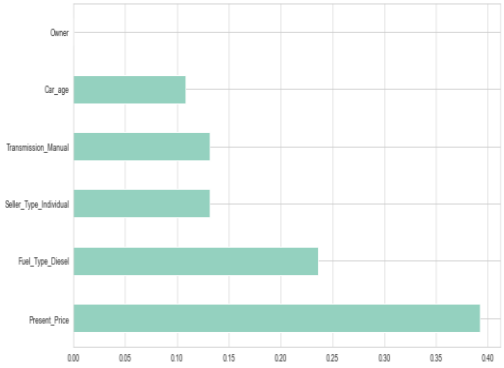


Another characteristic with a distinct subcategory in the used automobile market is the transmission. People's preferences for cars are heavily influenced by automatic gearbox. The global economic downturn might harm the used automobile industry. Another intriguing tendency is that other transmission has been increasing since 2009. Although it still has a small market share compared to automatic transmission, it is nonetheless significant. A number of factors might contribute to the rise of different transmission types. The impact of a car's condition on its price is the subject of this research. Figure 5 shows how 'condition' has a significant impact on the median automobile price. The condition values, on the other hand, have a high number of outliers, which is to be expected in such a large dataset.

Figure5: Axes Subplot



In the data, we can see that characteristics aren't very well associated. However, after log converting the 'Price' column, the correlation with a few characteristics rose, which is a positive thing.



CONCLUSION

Four distinct machine learning algorithms were employed to anticipate the price of used automobiles in this article. Due to the large number of characteristics that must be taken into account for a precise prediction, car price prediction may be a difficult undertaking. Data collection and preparation are two important steps in the prediction process. In this study, methods were developed to normalize, standardize, and clean data so that machine learning algorithms could avoid extraneous noise. As a result, the price attribute had to be divided into classes, each of which comprised a range of prices, although this clearly produced further mistakes. The study's biggest weakness is the small number of records that were used. To anticipate automobile costs in the future, we want to collect more data and apply more complex approaches such as artificial neural networks, fuzzy logic, and genetic algorithms.

REFERENCES

DOI: 10.5281/zenodo.6364937
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam

Proceedings of the National Conference on Emerging Computer Applications (NCECA)-2022
Vol.4, Issue.1

1. Sameerchand Pudaruth
http://ripublication.com/irph/ijict_spl/ijictv4n7spl_17.pdf

2. Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric

3. Pattabiraman Venkatasubbu, Mukkesh Ganesh Used_Cars_Price_Prediction_using_Supervised_Learning_Techniques

4. USED CAR PRICE PREDICTION
Praful Rane1, Deep Pandya2, Dhawal Kotak3

5. Predicting Used Car Prices
Kshitij Kumbar, Pranav Gadre and Varun Nayak

6. Machine Learning Techniques for Predicting Used Car Prices
predicting-used-car-prices-with-machine-learning-techniques

7. Richardson, M. S. (2009). Determinants of used car resale value.

8. Used cars database. (n.d.) Retrieved from: https://www.kaggle.com/orgesleka/used-carsdatabase.

9. Listiani M. 2009. Support Vector Regression Analysis for Price Prediction in a Car Leasing Application. Master Thesis. Hamburg University of Technology

10. OLX. (n.d.), Retrieved from: https://olx.ba.

DOI: 10.5281/zenodo.6364937
ISBN: 978-93-5607-317-3 @2022 MCA, Amal Jyothi College of Engineering Kanjirappally, Kottayam