Car Price Prediction

*Note: Sub-titles are not captured in Xplore and should not be used

Abhay Yadav
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab, India, 144411
abhay19022000@gmail.com

Anurag Singh Patel
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab, India, 144411
anuragsinghpatel98@gmail.com

Akash mor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab, India, 144411
akashmor7@gmail.com

Chavi Ralhan
Assistant Professor
School of Computer Science and Engineering
Lovely Professional University
Phagwara, Punjab, India, 144411
chavi.12086@lpu.co.in

1) Abstract— India has considerable size car sell on top of the world day-to-day. many buyers usually sell their cars after using for the time to another buyer, we name them as second possessor.numerous platforms such as carwale.com, cartrade.com, cars24.com, OLX.com and cardekho.com *etc. that come up with these buyers with a platform where they can sell their old cars, but what should be the price of the car, this is the long-lasting query ever by using Machine Learning algorithms we can lead a response to this issue. using a history of previous used car sales data and machine learning methodologies like Supervised Learning, I was able to predict a fair price for the car. I also used machine learning techniques like Random Forest and Extra Tree Regression, as well as the popular Python package Scikit-Learn.*

To estimate the old car's resale value The outcome has shown that these two algorithms are extremely accurate in prediction even the dataset is huge or small, irrespective of the size of the dataset they give an exact result.

Keywords—car price prediction, flask, machine learning, regression, interface

II. INTRODUCTION

Car price prediction is somehow interesting and popular problem. As per information that was gotten from IPSOS report currently, close to 5 million used cars are being sold in India every year, and millennials account for 80 percent of its sales. The used car market in the country is expected to reach over 7 million by 2015-26, according to a report by OLX Autos. The emergence of online portal such as car24, carfirst, cardekho, quikr, cartrade and many others has facilitated the need for both the customer and the seller to be better informed about the trends and patterns that determine the value of the used car in the market. Whenever a car is sold to dealers the price should be calculated. Prediction techniques of Machine Learning algorithms are used to predict the estimated price of used cars.

The data set used for the prediction models the following are the variable used.

1.Company: Specific Company name
2.Model: The Specific models for each car.
3. Year: Year of purchase car.
4. Fuel: type of car e.g., petrol, Diesel, CNG.
5. Number of Kilometers that car has travelled

III. RELATED WORK

Gonggie [6] suggested an ANN (Artificial Neural Networks)-based algorithm for predicting the price of a secondhand car. He took into account a number of factors, including the number of miles driven, the expected automobile life, and the brand. The proposed model was created to deal with nonlinear data relationships, which was not possible with previous models that used simple linear regression approaches. Other linear models were unable to predict automobile prices with the same precision as the non-linear model.

Richardson proposed a different strategy in his thesis [3]. His theory was that vehicle manufacturers make cars that are more robust. Richardson used multiple regression analysis to show that hybrid cars hold their worth for a longer period of time than standard vehicles. This stems from environmental worries about climate change, and also improves fuel efficiency.

In addition, Pudaruth [7] used a variety of machine learning methods for car price prediction in Mauritius, including k-nearest neighbors, multiple linear regression analysis, decision trees, and nave bayes. The data used to build the prediction model was personally gathered from local newspapers over a period of less than one month, as time can have a significant impact on car prices. He looked at brand, model, cubic capacity, kilometers per gallon, manufacturing year, exterior color, transmission type, and price. However, the author discovered that Naive Bayes and Decision Tree were incapable of classifying and predicting numeric values. Furthermore, because to the low amount of dataset instances, high classification performance, i.e., accuracies less than 70%, was not possible.

IV. METHODOLOGIES

A- **Data Processing:**

"Prior to Training, Data Preprocessing is that very important step and will be the first step. Data Preprocessing contains a number of steps such as: "

Step1: Import Libraries: Essential Libraries for Predicting Data I have used Pandas for data manipulation and analysis, NumPy numerical analysis, Matplotlib and Seaborn for better visuals and data image statistics.

Step 2: Import Database: This downloaded the database from Kaggle.

Step 3: Managing Lost Data on Data Set: After checking this database, I did not find any missing data in the database.

Step 4: Dividing the Data Set into the Training Set and Test Set: Dividing this database into an Exam and Train Database to train our machine learning model using a python, scikit-learn or sklearn machine library. It uses its model selection method to create test data by selecting random values from the available data for predictive modeling, or we might say Supervised Reading.

Step 5: Feature Scaling: Since all the data, available in a standard format, so here I do not use any feature scaling techniques

B- Model Comparison and Evaluation:

Variable values for price, fuel type, kilometers travelled, model, company, and model year, among other things, were collected. All attributes were considered at first, but we later used variable selection techniques on our data and eliminated all insignificant variables.

Figure 1 shows a sample of the data we used to feed into our price prediction regression model.

	Model	Company	year	Price	Kms_Driven	Fuel_Type
2	Hyundai Santro Xing XO eRLX Euro II	Hyundai	2007	80000	45000	Petrol
3	Mahindra Jeep CL550 MDI	Mahindra	2006	425000	40	Diesel
4	Hyundai Grand i10 Magna 1.2 Kappa	Hyundai	2014	325000	28000	Petrol
5	Ford EcoSport Titanium 1.5L TDCI	Ford	2014	575000	36000	Diesel
6	Ford Figo	Ford	2012	175000	41000	Diesel
7	Hyundai Eon	Hyundai	2013	190000	25000	Petrol
8	Ford EcoSport Ambiente 1.5L TDCI	Ford	2016	830000	24530	Diesel
9	Maruti Suzuki Alto K10 VXI AMT	Maruti	2015	250000	60000	Petrol
10	Skoda Fabia Classic 1.2 MPI	Skoda	2010	182000	60000	Petrol
11	Maruti Suzuki Stingray VXI	Maruti	2015	315000	30000	Petrol
12	Hyundai Elite i20 Magna 1.2	Hyundai	2014	415000	32000	Petrol

Fig.1. Dataset used for training

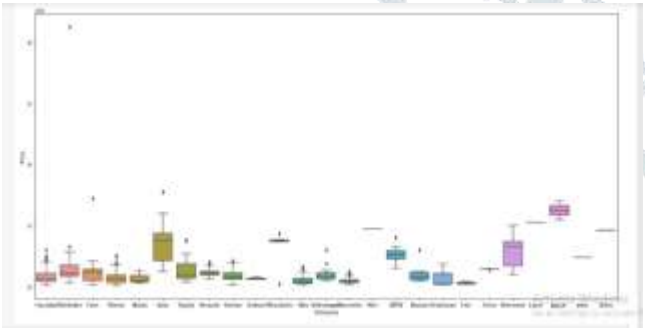

Fig.2. Company wise average prediction

Above chart shows a distinguished pricing amongst each company which proves that this feature would be important for prediction.

Now, we can use a few regression models to compare the accuracy of each model. For each model, we use 5 folds cross-validation.

It will divide the data randomly into 5 folds and trains the model 5 times. All fold will be selected for testing each time while the other 4 will be kept for training.

Model we are going to use Multiple Regression, Linear Regression and Decision tree.

Metrics are considered to test the strength of the algorithm by Mean, Standard deviation. These two metrics are used to test the three algorithms mentioned above. Two metrics have low Multiple regression values, which is why it has a higher accuracy than the other two algorithms. Table 1 shows the analysis of the results of all two algorithms.

Table 1. ANALYSIS OF ALGORITHM

Algorithm	Mean	Standard deviation
Decision Tree	4.189504502474483	0.848096620323756
Multiple Regression	3.494650261111624	0.762041223886678

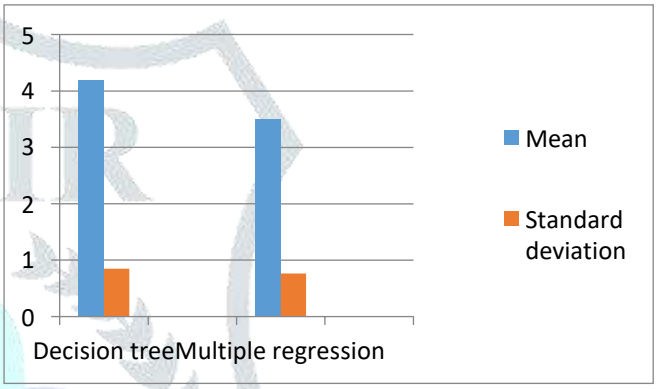

Fig.3 Analysis of algorithms

Multiple regression is an improvised version of linear regression model where it helps in fitting relationship between more than two variables. Mathematical equation of multiple regression model is:

$$Y=\beta0 + \beta1X1 + \beta2X2 + \ldots\ldots + \beta nXn$$

Here, Y is the dependent variable and X1, X2, ….Xn are the independent variable. β0, β1, …., βn are the regression coefficients of the independent variables X1, X2,…,Xn.. Use of independent variables helps in determining the values of regression coefficients by fitting the model. This model is used to identify the dependent variable.

The predictions may not always be accurate. There will be miscalculations. Figure 3 contains actual and predicted price in fit column found in dependent variables.

Price	Fit
1560000	1372547
1090000	1277393
850000	850000
1000000	984842
907750	938065
875000	1065347
800000	800000
2850000	2850000
1750000	1750000
5200000	5200000
2650000	2575000

Fig.4. Actual vs fitted price

Price prediction can be done by giving input values of the required information, i.e., company, model, model year, kms driven and fuel type in the proposed regression model.

V. CONCLUSION

The new car market is now expensive and not everyone can afford it so the used car market will be the future for everyone can afford it at a very low cost but there is a problem with the market that the total price of the used car is determined with the seller can also set any price as he pleases but this creates a problem in the market. The proposed system works well to predict the fair value of pre-owned vehicles. The system wisely operates the prices of used cars. System user be it a dealer or a buyer, you will get a reliable amount of used car.

REFERENCES

[1] GONGGI, S., 2011. New model for residual value prediction of used cars based on BP neural network and non-linear curve fit. In: Proceedings of the 3rd IEEE International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Vol 2. pp. 682-685, IEEE Computer Society, Washington DC, USA.

[2] RICHARDSON, M., 2009. Determinants of Used Car Resale Value. Thesis(BSc). The Colorado College.

[3] S. Pudaruth, "Predicting the Price of Used Cars using Machine Learning Techniques," International Journal of Information & Computation Technology, vol. 4, no. 7, pp. 753–764, 2014.

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

[5] WU, J. D., HSU, C. C. AND CHEN, H. C., 2009. An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference. Expert Systems with Applications. Vol. 36, Issue 4, pp. 7809-7817.

[6] DU, J., XIE, L. AND SCHROEDER S., 2009. Practice Prize Paper - PIN Optimal Distribution of Auction Vehicles System: Applying Price Forecasting, Elasticity Estimation and Genetic Algorithms to Used-Vehicle Distribution. Marketing Science, Vol. 28, Issue 4, pp. 637-644.

[7] S. Peerun, N. H. Chummun, and S. Pudaruth, "Predicting the Price of Second-hand Cars using Artificial Neural Networks," The Second International Conference on Data Mining, Internet Computing, and Big Data, no. August, pp. 17–21, 2015.

[8] N.Sun, H. Bai, Y. Geng, and H. Shi, "Price evaluation model in second-hand car system based on BP neural network theory," in 2017 18th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), jun 2017, pp. 431–436.

[9] Geurts P. (2009) Bias vs Variance Decomposition for Regression and Classification. In: Maimon O., Rokach L. (eds) Data Mining and Knowledge Discovery Handbook. Springer, Boston, MA

[10] Robert T. (1996) Regression Shrinkage and Selection Via the Lasso. In: Journal of the Royal Statistical Society: Series B (Methodological) Volume 58, Issue 1.