Prediction of Used Car Price Based on Supervised Learning Algorithm

Feng Wang
College of Computer Studies
Angeles University Foundation
Angeles, Philippines
Wang.Feng@auf.edu.ph

Xusong Zhang*
College of Computer Studies
Angeles University Foundation
Angeles, Philippines
Zhang.xusong@auf.edu.ph

Qiang Wang
College of Computer Studies
Angeles University Foundation
Angeles, Philippines
Wang.quiang@auf.edu.ph

***Abstract*—In this paper, we use machine learning algorithms to predict the price of used cars with less human intervention to make the results more objective. The method used is to preprocess the dataset through Python's Pycaret package and compare the performance of each algorithm through the algorithm comparison function, in this study Extra Trees Regressor, Random Forest Regressor performs relatively well. Finally, the algorithm was optimized by using the hyperparameter function. The results show that R2 = 0.9807 obtained from extreme random numbers is the best performance. The algorithm was obtained and validated with new data to derive the final algorithm model. When new used car data flows into the used car system, used car prices will be automatically generated by this algorithm, which will make the workflow of the used car market faster and more competitive for that used car market.**

***Keywords*—machine learning, supervised learning, used car price, prediction**

I. Introduction

According to relevant reports, in the next five years, the annual growth rate of automobiles in China will be 3.5%, while the annual growth rate of used cars will be 5%. The annual growth rate of used cars and automobiles is constantly expanding. Therefore, consumers think that when buying a new car, they will also consider the price of the same type of used car, especially some value-preserving brand cars are more worthy of consumers' attention, which is a change of value, and consumers can get the best return on investment. Faced with this situation, companies operating the used car market use traditional marketing methods (consulting prices for many times) to deal with business, which greatly increases the company's operating costs. This paper will predict the used car prices through various supervised learning algorithms in machine learning, and the used car companies can directly publish the predicted prices through Internet channels, so that consumers can know the used car prices at a glance and provide operational efficiency of the company.

II. Literature Reviews

Looking at the global research on the price prediction of used cars, it is found that many experts and scholars have done research, such as forecasting the price of used cars by linear regression, Bayesian, decision tree and other algorithms, and forecasting the price of used cars by neural networks. For example, the training samples are defective, the number of samples is not enough, and all the relevant mainstream algorithms have not been trained and compared [1]. If only several related algorithms are selected for prediction by subjective judgment, the conclusion is not sufficient. In this paper, when selecting data sets, we will use large data samples for training, verify the performance through independent test samples, predict through all mainstream algorithms in supervised learning, sort the performance from high to low, select 1-2 optimal algorithms for in-depth analysis, and finally get the prediction model, which will be tested with test samples [2].

III. Methodology

This section will be described in sequence according to the figure 1:

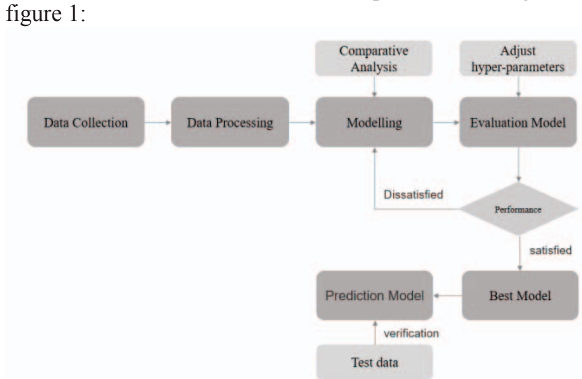

Figure. 1. Process of this forecast model

A. Source and Description of Data Set

The data set of this study comes from the second-hand car price data set in Ali Tian chi from 1996 to 2019(download:https://tianchi.aliyun.com/dataset/dataDetail?dataId=93470), with a total of 6019 data items. In order to prevent information leakage, the data set is randomly divided into training set and test set according to the ratio of 8:2. Including 12 feature items such as name, location, year, kilometers _ driven, fuel _ type, transmission, owner _ type, mileage, power, seats, new _ price, price, etc., among which price needs to be predicted, and the specific meaning of each feature item is shown in Table II [3].

The features and labels of the data set are described as the table I:

TABLE I DATA SET FEATURES AND LABEL DESCRIPTION.

Attribute	Description
Name	The brand and model of the car
Location	The location in which the car is being sold or is available for purchase
Year	The year or edition of the model

Kilometers_Driven	The total kilometers driven in the car by the previous owner(s) in KM
Fuel_Type	The type of fuel used by the car
Transmission	The type of transmission used by the car
Owner_Type	Whether the ownership is Firsthand, Second hand or other
Mileage	The standard mileage offered by the car company in kmpl or km/kg
Engine	The displacement volume of the engine in cc
Power	The maximum power of the engine in bhp
Seats	The number of seats in the car
New_Price	Price of new model
Price	The price of the used car in INR Lakhs

B. Data Set Preprocessing

The data set is randomly divided into training set and test set according to the ratio of 8: 2.

TABLE II TRAINING SET AND TEST SET ARE RANDOMLY DIVIDED ACCORDING TO 8:2.

	…	Year	Kilometers_Driven	Fuel_Type	…	Price
0	…	2010	72000	CNG	…	1.75
1	…	2015	41000	Diesel	…	12.50
2	…	2011	46000	Petrol	…	4.5
3	…	2012	87000	Diesel	…	6.0
…	…	…	…	…	…	..

Date for Modeling: (4815,13)
Unseen Date For Predictions: (1204,13)

As shown in table III. In actual business processing, data is usually dirty. Dirty means that the data may have the following problems (main problems) [4]:

1. Missing data is when the attribute value is empty. E.g. Occupancy = ""

2. Data noise is an unreasonable data value. E.g. Salary = "-100 "

3. Inconsistency of data means that there are contradictions before and after the data. E.g. age = "42" vs. birthday = "01/09/1985"

4. Data redundancy refers to the situation that the amount of data or the number of attributes exceeds the need of data analysis.

5. Unbalanced data sets are situations in which the amount of data in various categories varies greatly.

6. Outliers/outliers are data far away from the rest of the data set.

7. Data duplication is data that appears many times in a data set.

TABLE III MISSING FEATURE DATA OF DATA SET.

Name	Location	Year	…	Power	New_Price	Price
Honda City 1.5 GXI	Ahmedabad	2007	…			2.95
Land Rover Range Rover 3.0 D	Mumbai	2008	…			26.5
Honda City 1.3 DX	Delhi	2009	..			3.2
…	…	…	…			…

Measures to deal with missing data.as shown in the table IV:

1. Delete data, and if the proportion of records with missing data is small, delete these records directly. (unbalanced data, discrete data, repeated data, etc.).

2. Fill in manually, or collect data again, or supplement data according to domain knowledge. (lack of text data)

3. Automatic filling: mean filling (missing numerical data).

4. Text data conversion.

After filling the data set completely, it is necessary to encode the data.

Among the 12 features, 10 are classified features, and 2 are numerical features. The features are normalized, and then each feature is transformed by scaling the quartile distance on the basis of normalization. This transformation will have better results when the data set contains outliers.

TABLE IV DATA NORMALIZATION

Name_AudiName Audi somo di.. Name_Audi	Name_AudiA4 2.0 TDI celebration edition	Name_Audi.udi177 Bhp A42.0iDA4 2.0 TDI Name AtaA4 3.0 TDI	Name_Audia4.3.0 TDI quattro	name_CelebrationMltitroncA4 2.0 TFSl
0.0	0.0	0.0	0.0	0.0
….	….	….	…	…
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
…	…	…	…	…

C. Create Models and Compared

By comparing the model functions, it is concluded that the regression algorithm is sorted from big to small according to R2.as shown in table V.

TABLE V PERFORMANCE COMPARISON OF REGRESSION ALGORITHMS.

	MODEL	MAE	MSE	RMSE	R2	RMSLE	MAPE
et	EXTRA TREES REGRESSOR	1.7464	17.7966	4.104	0.8614	0.2327	0.1944

rf	RANDOM FOREST REGRESSOR	1.9045	20.1534	4.3821	0.8448	0.2424	0.2103
ridge	RIDGE REGRESSION	2.2622	20.3384	4.43	0.8417	0.3425	0.4104
dt	DECISION TREE REGRESSOR	2.2628	25.9082	4.9663	0.8001	0.3042	0.2606

RMSE (square root error), MAE (mean absolute error), MSE (mean square error), R2, TT(sec), MAPE, RMSLE (root mean square log error), etc. are all performance indicators. when the dimensions are different, R2 is relatively more accurate. When R2 is closer to 1, the regression fitting effect is better. In the above figure, ET (Extra Trees Regression) R2 is 0.8614, and other related performance indicators such as MAE, MSE, RMSLE and MAPE are the best, which shows the best performance. (yellow mark) [4].

ET implements a meta-estimator, which fits many random decision trees (additional trees) on various subsamples of the data set, and uses averaging to improve prediction accuracy and control over-fitting. He is the same as the random forest in other aspects except the number of training samples and bifurcation mechanism, so his performance is close to that of the random forest [5].

After ET is constructed, we can also apply all the training samples to get the prediction error of ET. This is because although the same training sample set is used for building decision tree and forecasting, because the best bifurcation attribute is randomly selected, we will still get completely different forecasting results, which can be compared with the real response value of samples, thus obtaining the forecasting error. If compared with random forest, all training samples in ET are OOB samples, so calculating the prediction error of ET is to calculate this OOB error [6].

Firstly, the ET algorithm is modeled, and 10-fold cross-validation is adopted to obtain the following data. As show in table VI.

TABLE VI ET ALGORITHM 10 TIMES CROSS VALIDATION.

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	1.4749	10.8751	3.2977	0.8622	0.2471	02289
1	1.5452	11.3756	3.3728	0.8907	0.2086	0.1830
2	2.0896	18.8049	4.3365	0.8759	0.2338	0.1927
…	…	…	…	…	…	…
9	1.6551	10.5013	3.2406	0.9047	0.2226	0.1890
Mean	1.7464	17.7966	4.1040	0.8614	0.2327	0.1944
SD	0.2414	9.1411	0.9768	0.0478	0.0170	0.0167

The evaluation model of ET is verified, and the verification curve and learning curve are obtained:

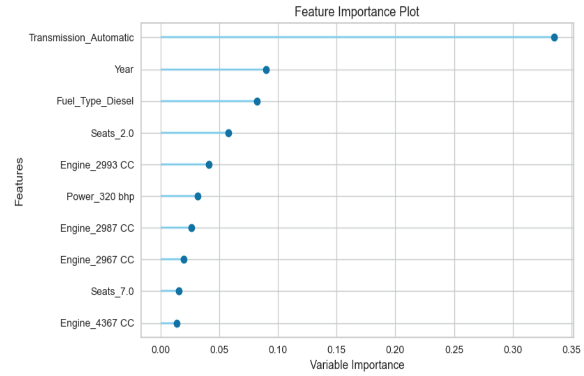

Figure. 2. Feature importance plot

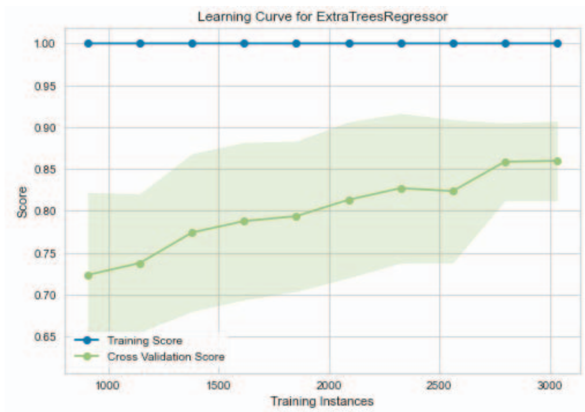

Figure. 3. Learning curve

According to the figure 2 and figure 3, it is found that the normalized feature Transmission_automatic is the most important and has the greatest influence on the algorithm. Others are year fuel _ type _ diesel, seats _ 2.0, engine _ 2993cc, power _ 320bhp, engine _ 2987cc, engin _ 2967cc, seats _ 7.0, engine _ 4367cc, etc. other features can be deleted to increase the speed of the algorithm. The results are as follows:

TABLE VII ET PERFORMANCE WITH 15 ITERATIONS.

	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	1.8937	18.5397	4.3058	0.8456	0.2471	0.2147
1	2.2102	21.6639	4.6545	0.8679	0.2482	0.2084
2	1.8048	13.9492	3.7349	0.8642	0.2395	0.2145
…	…	…	…	…	…	…
9	1.7405	13.3783	3.6576	0.8786	0.2421	0.1933
Mean	1.7375	14.5607	3.7689	0.8787	0.2330	0.2011
SD	0.2039	4.5866	0.5967	0.0346	0.0203	0.0159

After 15 iterations, the average R2 of ET increased from 0.8614 to 0.8787. as show in table VII

D. Generating a Prediction Model.

TABLE VIII PREDICTION PERFORMANCE

	Model	MAE	MSE	RMSE	R2	RMSLE	MAPE
0	Extra Trees Regressor	1.9056	28.6677	5.3542	0.7862	0.2444	0.1954

TABLE IX PREDICTION RESULTS

New_price_9.72 lakh	….	New_price_not_available	Price	Label
0.0	….	0.0	5.85	6.7689
0.0	….	0.0	4.94	5.2829
…	…	…	…	…
0.0	…	0.0	1.00	1.0535
…	…	…	…	…

As shown in table VIII and table IX. The R2=0.7862 (the value in the red circle) generated by the predict_model(et) function is far from the previous R2=0.8787, so there may be over-fitting. 20% of the reserved test data is predicted. the label value in the green box means that there is a certain gap between the predicted value and the price value. after fitting this data with the finalize_model(et) function, R2 = 0.9807 is obtained. the finalize _ model (et) function is used to predict the whole data set, while the previous predict_model(et) function is only for the data of the training set. It can be seen that 1445 test set data are very important for the whole prediction model [7].

TABLE X FORECAST RESULTS FOR DATA SETS WITH 20% RESERVATION.

	Name	Location	Year	…	Price	Label
0	Tata Sumo Ex	COIMBATORE	2015		5.29	5.29
1	Eatsun redi-GO T	KOLKATA	2016		2.25	2.25
…	…	…	…	…	…	..

It can be seen from the table X that the predicted value is basically consistent with the actual price.

IV. RESULTS

Through the comparison of a series of algorithms of supervised learning regression algorithm, it is concluded that et algorithm has the best performance [8]. Through data preprocessing, super-parameter adjustment and other operations, R2=0.9807 is finally obtained, and 20% of the reserved data is tested to meet the expected standard [9]. Next, we can automatically get the predicted price of used cars by et algorithm after collecting all the 11 feature values on the Internet. As shown in the table XI [10].

TABLE XI NEW DATA PREDICTION RESULTS

	Name	Location	Year	…	Power	New_Price	Label
0	Maruti Alto K10 LXI CNG	DELHI	2014	…	58.2 bhp	NaN	2.5788
1	Maruti Alto 800 2016-2019 LXI	Coimbatore	2013	…	796cc	NaN	2.8466
2	Toyota Innova Crysta Touring Sport 2.4 MT	Mumbai	2017	…	147.8bhp	25.27Lakh	20.7864
3	Toyota Etios Liva GD	Hyerabad	2012	…	Null bhp	NaN	4.0455
…	..	…	…	…	..	..	..
1232	Volkswagen Polo GT TSI	Pune	2013	…	103.6bhp	NaN	4.4217
….	…	…	…	…	…	…	…

The price of new used cars on the Internet platform is obtained by ET algorithm, and consumers can know the price directly by looking at the data.

V. CONCLUSION

In this study, the author makes a series of performance comparisons based on supervised learning algorithms. The data set used here comes from the price of used cars, and python language is used to predict the data set [11]. It can be seen from the results that we compare the performance by using several algorithms, such as ET, rf, ridge, and so on. Each model is tested by using the same training data. The results are compared with the average absolute error and further demonstrated by the multi-dimensional evaluation model.

Then the best performance model is selected as the prediction model, and finally verified by the new used car data. The result given from the best performing algorithm model is R2=0.9807, and the final verification of new data shows that ET alalgorithm is the best model for the second-hand car price prediction, and it will be more in line with the daily operation by adjusting the super parameters in the future work [11].

In practical application, inputting all kinds of characteristic data through the Internet port will directly display the prediction results on the port interface, which greatly improves the working efficiency of the used car market, thus improving its market competitiveness.

REFERENCES

[1] C. Chen, L. Hao, and C. Xu, "Comparative analysis of used car price evaluation models", 2017 pp. Pages.

[2] N. Monburinon, P. Chertchom, T. Kaewkiriya, S. Rungpheung, S. Buya, and P. Boonpou, "Prediction_of_prices_for_used_car_by_using_regression_models", *2018 5th International Conference on Business and Industrial Research (ICBIR),* 2018, pp. 115-119.

[3] 1*, I.S., 1*, M.Z., and 2*, S.: "Analysis of The Application of Fuzzy Logic and Levenberg-Marquardt in The Calculation of Used Car Prices", *ICCAI 2019 Journal of Physics: Conference Series*, 2020.

[4] C. V. Narayana, C. L. Likhitha, S. Bademiya, and K. Kusumanjali, "Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business". P*roc. 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC),* 2021, pp. Pages.

[5] Q. Chen, and S. Lee, "A Machine Learning Approach to Predict Customer Usage of a Home Workout Platform", *Applied Sciences*, 2021, 11, (21).

[6] V. Bhadana, A. S. Jalal, and P. Pathak, "A Comparative Study of Machine Learning Models for COVID-19 prediction in India", in Editor (Ed.)^(Eds.): 'Book A Comparative Study of Machine Learning Models for COVID-19 prediction in India' (IEEE, 2020, edn.), pp. 1-7.

[7] S. K. Budhani, R. S. Bisht, and N. Budhani, "A Model for prediction of consumer conduct using machine learning algorithm", in Editor (Ed.)^(Eds.): 'Book A Model for prediction of consumer conduct using machine learning algorithm' (IEEE, 2020, edn.), pp. 256-260.

[8] D. Dansana, J. D. Adhikari, M. Mohapatra, and S. Sahoo, "An approach to analyse and Forecast Social media data using Machine Learning and Data Analysis", in Editor (Ed.)^(Eds.): 'Book An approach to analyse and Forecast Social media data using Machine Learning and Data Analysis' (IEEE, 2020, edn.), pp. 1-5.

[9] H. Jiang, J. Ruan, and J. Sun, "Application of Machine Learning Model and Hybrid Model in Retail Sales Forecast". Proc. 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA) 2021 pp. Pages.

[10] D. Mishra, and P. Joshi, "A Comprehensive Study on Weather Forecasting using Machine Learning". Proc. 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO) 2021 pp. Pages.

[11] H. Lei, and H. Cailan, "Comparison of Multiple Machine Learning Models Based on Enterprise Revenue Forecasting". Proc. 2021 Asia-Pacific Conference on Communications Technology and Computer Science (ACCTCS) 2021 pp. Pages.