Used Car Prices in India: What about Future?

Xinru Chen[1],Shiyu Gu[2]*, Ximing Deng[3], Lei Huang[4]

[1]*University of Washington, USA*
[2]*University of Sydney Business School, Australia*
[3]*Gies College of Business, University of Illinois,Urbana-Champaign, USA*
[4]*Department of Physics, University of Washington, USA*
[1]*e-mail: chenxr@uw.edu,*
[*]*e-mail: shgu4120@uni.sydney.edu.au*
[3]*e-mail: ximingd2@illinois.edu*
[4]*e-mail: leih3@uw.edu*
These authors contributed equally.

ABSTRACT

With the steady increase in the demand for cars and the shrinking supply of the new cars due to chip shortage in the worldwide market, the used car price has risen dramatically in recent years. This paper investigated Used Car prices in India market by applying multiple machine learning models to predict and analyze the typical characteristics of the used car market in India, which is a rapidly developing country that has an ill-informed used vehicle buyer market. Before building models, We checked important statistical attributes of our data, created a new variable called the average_cost_price, converted the unit of price to thousand dollars, visualized the data through different graphs, and dropped several potential outliers to eliminate the effects on the later models. Then, we applied four useful machine learning algorithms, including Linear Regression, Decision Tree Regression, Random Forest Regression and Gradient Boosting Regression, for model buildings and analysis. Based on the performance of our final optimal model, the final optimal Gradient Boosting Model can be used to predict used car prices in India market at the current point. Buyers can then get much more information about used car prices in India, and sellers are hard to hide information in face of our relatively comprehensive model.

Keywords*-Machine Learning, Gradient Boosting, Decision Tree, Used Car, Information Asymmetry*

1. INTRODUCTION

According to the data reported from the market research by the Hedges Company [1], there are about 1.446 billion vehicles worldwide in 2021. Many people would think North America is the place that is most preoccupied with cars. However, over one-third of the total number of vehicles are from Asia, and then Europe ranks second. Including trucks, there are 518 million vehicles on the road in Asia. Judging from the current situation (OECD predicts growth to reach 5.7% this year and 4.5% in 2022 worldwide, i.e., the economic strength of people worldwide keeping ascending), this figure is bound to continue to grow.

Figure 1. Used Car Market-Growth Rate. *Notes:* From Mordor Intelligence.

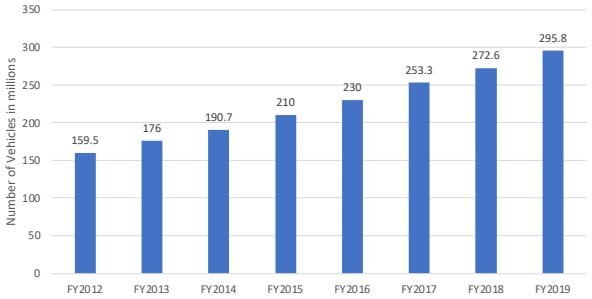

Figure 2. Number of registered Vehicles across India from Financial Year 2012 to 2019.

Fig. 2 shows that the number of registered vehicles steadily increases and continuing to grow at an increasing rate, and as the growth of the Cars in the world per capita by world region, the used car market is also the one that cannot be ignored. According to the market research report by Mordor Intelligence [2], the used car market is expected to reach USD 403 billion by 2026. Registering a CARG (Compounded Average Growth Rate) of more than 10% during 2021 and 2026. Meanwhile, Mordor Intelligence's report introduced that the market is also being driven by rising income levels, shrinking car ownership periods, booming import and export trade, and a growing trend among people to upgrade to smaller and compact vehicles. The Covid-19 pandemic also unavoidably affects the market. Beginning in February 2020, sales automotive industry has dropped 71 percent in China, 47 percent in the US, and 80 percent in Europe [3]. The issue of shortage of available new cars has become more serious under the pandemic, as the increase in car demand never stops. Therefore, we should pay more attention to the used car market.

Considering the importance of the used car market and the unassailable proportion in the vehicle market of Asia explained above, we decided to dig deeper into the used car markets in a developing country India. The India used-car market was valued at USD 27 billion in 2020, and it is expected to reach USD 50 billion by 2026, registering a CAGR of 15% during 2021 and 2026 [2]. Building and examining different predictive models in a typical market as in India will provide us with valuable insights and implications for our research problem.

Meanwhile, the used car market has long been considered a 'lemon' market [5]. Information asymmetry between buyers and sellers leads to buyers being unable to obtain sufficient vehicle information, and sellers can take advantage to sell lower-quality cars to ill-informed buyers. Akerlof warned that the market may crash in the limit, because the average quality of cars will be lower than that of the population of cars in the 'lemon' used car market, thus low-quality cars will drive high-quality cars out of the market [4]. One possible way to solve the harm dealt by information asymmetry in the used car industry is to set up a blockchain platform that would make information about second-hand cars more transparent [6]. However, that would take years to set up, therefore since the information asymmetry is a serious issue in the second-hand car market, and the price is one of the most important factors acting on the decision of a car purchase, it is crucial to predict the price as it would effectively decrease the market failure caused by adverse selection and moral hazard, which will increase consumption utility.

This paper aims to predict the price trend of second-hand cars in India, and study car prices in the highly promising Indian used car market and help solve the problem posed by the lemon phenomenon, the dataset we used contains more than 15000 used car information from cardehko.com, and 13 potential factors that may have effects on the car price. These large amounts of data information can solve the problem of information asymmetry to a certain extent and provide buyers with transparent market information to predict the price of vehicles, to maintain the used car market.

This study uses several popular machine learning techniques to analyze the dataset. Three performance metrics (Root Mean Squared Error, Mean Absolute Percentage Error and Coefficient of determination) are used to compare the fitting degree and error size of the above four models.

Importantly, we give a potential way to solve information asymmetry by analyzing a dataset with machine learning methods, which will significantly increase the consumption utility of second-hand cars in India and provide consumers with a fairer market.

2. Data and Variables

2.1. Data Description

The dataset we used was gathered by Sai Saathvik in 2021, from Kaggle.com. This dataset contains more than 15,000 sets of important attributes information of used cars sold on cardehko.com in India. The dataset consists of 10 numerical variables, including min cost price, max cost price, vehicle age, km driven, mileage, max power, seats, engine, and selling price; and 6 categorical variables, including seller type, fuel type, transmission type, car name, brand, model. Table 1 provides a summary of all attributes of the 6 categorical variables, and Table 2 provides a summary of the important statistical features of the 9 numerical variables.

TABLE I. SUMMARY OF 6 CATEGORICAL VARIABLES

Variables	Categories
car_name	Maruti Alto, Hyundai Grand, Isuzu, etc.
brand	Maruti, Hyundai, Honda, BMW, etc.
model	Represents the specific car model from the above car brands
seller_type	Individual Dealers, Trustmark Dealers.
fuel_type	Petrol, Diesel, LPG, CNG, Electric
transmission_type	Automatic, Manual

TABLE II. STATISTICAL FEATURES OF 9 NUMERICAL VARIABLES

Variable name	count	mean	std	min	25%	50%	75%	max
Min_ cost_ price	15,411	5,892,971	56,698,510	311,000	668,000	855,000	1,284,000	988,000,000
Max_cost_price	15,411	1962,624	2,889,636	436,000	872,866	1,206,000	1,743,000	91,100,000
Vehicle age	15,411	6.04	3.01	0.00	4.00	6.00	8.00	29.00
Km_driven	15,411	55,616.48	51,618.55	100	3,000	50,000	70,000	3,800,000
Mileage	15,411	19.70	4.17	4.00	17.00	19.67	22.70	33.54
Engine	15,411	1,486.06	521.11	793	1197	1248	1,582	6,592

2.2. Selling Price

Selling price is the dependent variable in our analysis. Table 3 shows eight important statistical data of the selling price variable, including percentile, maximum and minimum value, and standard deviation. The table shows that the selling price of the used car in India ranged from $40.00 to $39,500.00 thousand. The mean is approximately 39% greater than the median value, which leads to a highly right-skewed distribution. Fig. 3. demonstrates that more data is clustered to the left portion of the graph which indicated that a substantial amount of secondary cars were sold at a price less than $5,000 thousand.

TABLE III. STATISTICAL FEATURES OF THE DEPENDENT VARIABLE

	Count	Mean
selling_price	15411	774971.10
	std	min
	894128.40	40000.00
	25%	50%
	385,000	556,000
	75%	max
	825000.00	39500000

Figure 3. Distribution Plot for selling_price

Fig. 4 is a box plot that displays the distribution of the selling price for each of the car brands. Several brands, such as Ferrari, Maserati, Mercedes-AMG, which are typically considered luxury cars contain only little data (less than 10). Therefore, the data of such car brands were considered as outliers and were dropped.

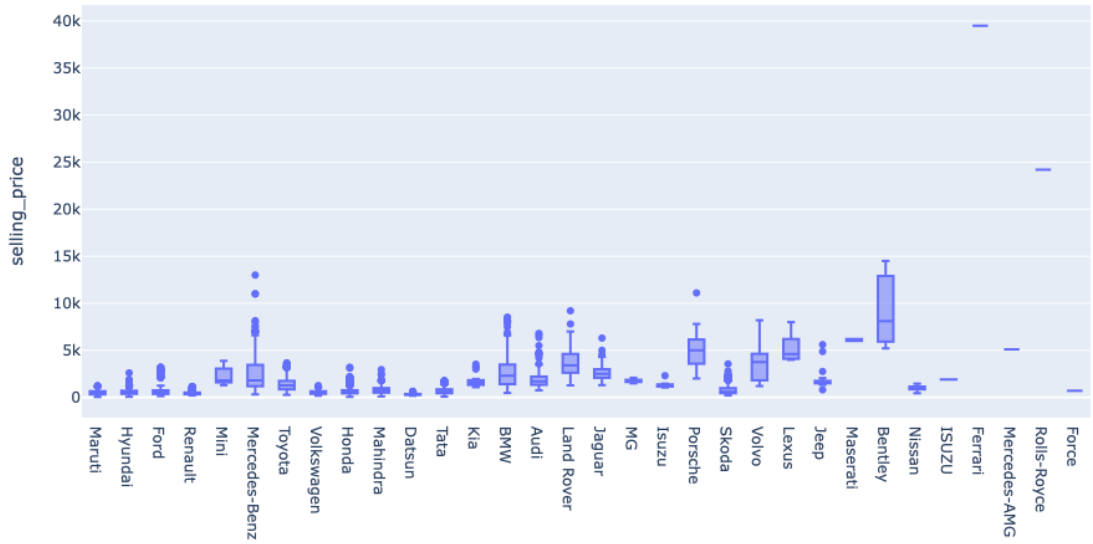

Figure 4. Box plot of selling_price

2.3. *Independent Variables*

Fig. 5 provides a summary of the histograms of all the numerical variables (the minimum and maximum cost price of the car were dropped, and the average cost price was calculated as one new numerical variable). The histogram shows the number of instances on the vertical axis that have a given value range on the horizontal axis. One notable point from these histograms is that most of the histograms are highly skewed, which may lead to some poor performances of certain models.

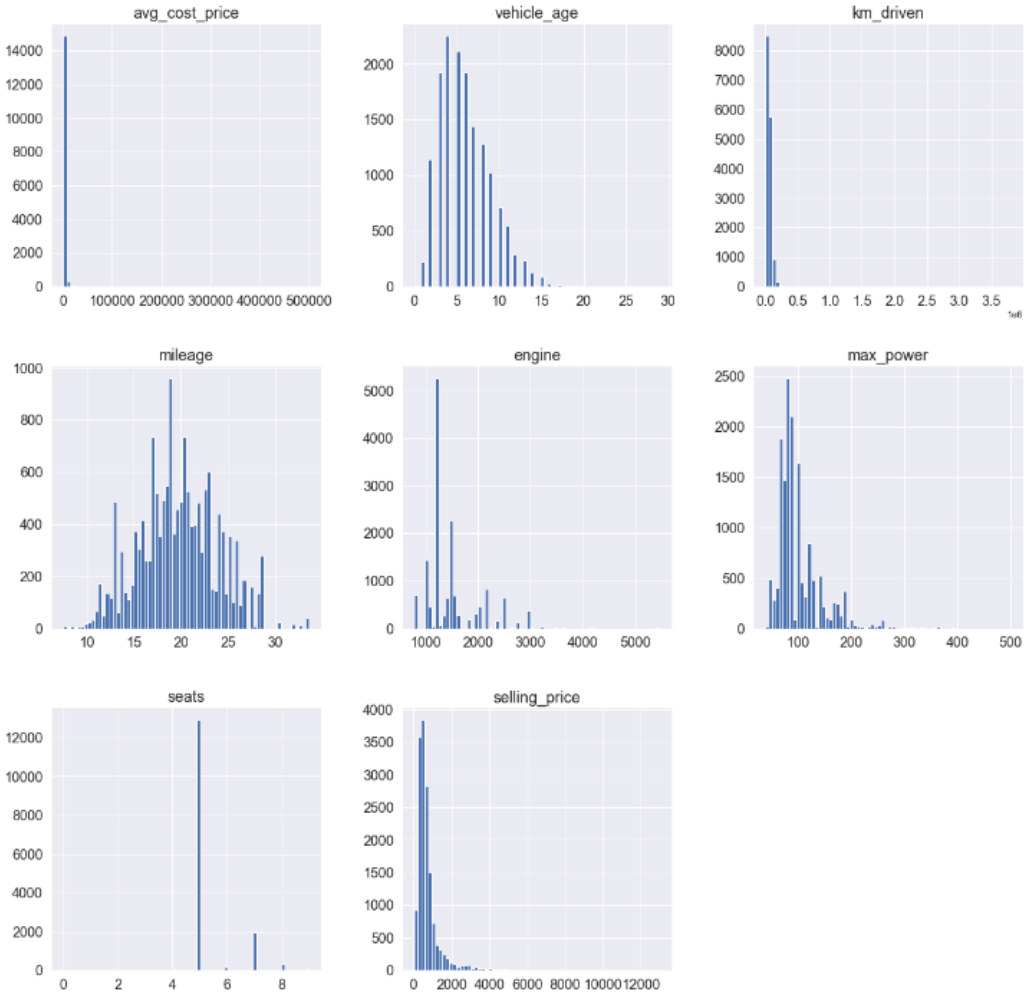

Figure 5. Summary of histograms of numerical variables

2.4. Correlations

Fig. 6 shows a Correlation Matrix of all the variables. Each cell in the matrix represents the correlation between two variables, and the values lie between -1 and +1. A negative number represents a negative correlation between the two variables, and a positive number represents a positive correlation. The further away the correlation coefficient is from zero, the stronger the relationship between the two variables.

Several notable points can be concluded from the correlation matrix. First, some of the predictors have higher correlations with selling_price, such as max_power_value and engine. Engine has the strongest correlation with the selling price of a used car. Meanwhile, other predictors such as km_driven and seats have nearly no correlation with selling_price, thus will not be considered when building models. Another unneglectable point is that some predictors are correlated to each other. For instance, engine and max_power have a strong correlation with a correlation coefficient value of 0.8. Therefore, the issue of collinearity will be carefully considered in the later model building process.

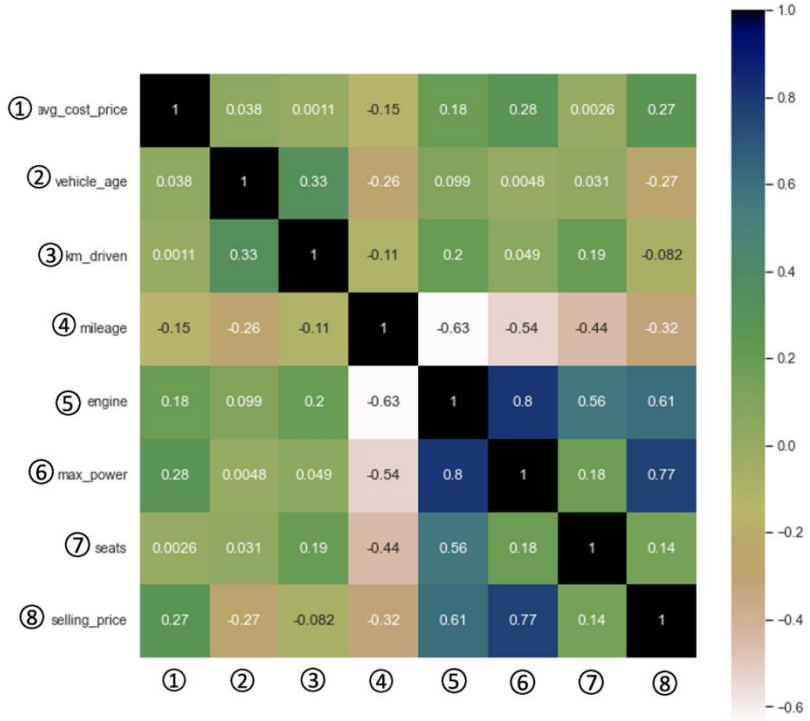

Figure 6. Correlation Matrix

3. METHODOLOGY

3.1. Algorithms

We applied the following four algorithms into our used car price valuation and analysis.

3.1.1. Linear Regression

Linear regression is a widely used method for modelling the relationship between a response variable (dependent variable) and one or more explanatory variables (independent variable). Linear regression can be divided into simple linear regression with a single independent variable and multiple linear regression with more than one independent variable [7]. The multiple linear regression will be applied in the case of modelling the secondary car price.

The general form of the multiple linear regression model is,

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon, \quad (1)$$

where y is a dependent variable, $\beta_0, \beta_1, \ldots, \beta_p$ are regression coefficients, and $x_1, x_2, \ldots, x_n$ are independent variables in the model. The error term ε follows the normal distribution with $E(\varepsilon) = 0$ and a constant variance $Var(\varepsilon) = \sigma^2$.

3.1.2. Decision Tree Regression

Decision tree is a common-used prediction model in data mining that is often used to treat large, complicated datasets with intelligibility and efficiency. In general, a decision tree model provides a prediction algorithm for a response variable by setting up classification systems based on different covariates [8]. This tree-structured classifier usually contains three types of nodes: Root Node, Interior Node, and Leaf Node. Each node in the tree represents an object, while each is split into paths according to different attributes. And each leaf node corresponds to the value of the attribute represented by the path from the root node. However, it should be noted

that the algorithm is binary which means that each non-leaf node can only split into two branches [9]. Thus, it can be interpreted as someone running through the entire tree by answering True/False questions to reach the final leaf node. Under multiple iterations, the Tree can analyze the data and give a proper prediction of the data point.

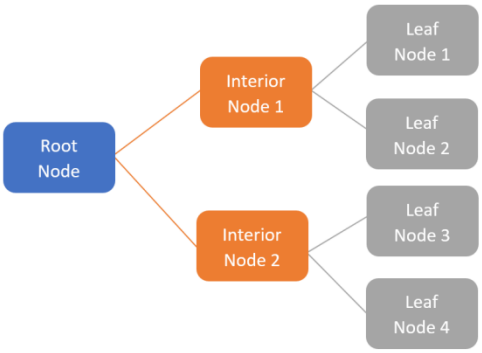

Figure 7. Decision Tree

3.1.3. Random Forest Regression

Decision tree regression algorithm is a machine learning tool with simplicity. However, a single tree may not be enough in most cases. Moreover, in the modelling of the decision tree, there is a risk of overfitting. Therefore, the Random Forest algorithm is often applied when facing more complicated requirements. The Random Forest is also known as a tree-based algorithm that contains features of multiple decision trees for decision makings. It uses the idea of ensemble learning to integrate multiple trees, and its essence belongs to the main branch of machine learning -- ensemble learning method. The term "random" comes from the creation of "randomly created decision trees". From an intuitive point of view, each tree is a classifier. Then, for an input sample, n trees will have n classification results. Random forest integrates all classification results and designates the category with the most votes as the final output, which is the simplest bagging idea in ensemble learning [10].

3.1.4. Gradient Boosting Regression

Boosting algorithm has a great effect on controlling bias and variance at the same time. It is a type of machine learning boosting with high efficiency. The key point of the gradient boosting algorithm is to set the target outcomes for the next model, combine the best possible next model with the previous models, and then minimize the gradient of the prediction error [11]. The target outcomes come from the changing of the case prediction's impact on the prediction error. In other words, the next target outcome has a high value when a small change in the prediction reduces the error greatly. Meanwhile, it has a zero value when a small change in the prediction does not cause any drop in error. In this way, every new model makes a progress on reducing the overall error [12].

The formula of loss in using $f(x)$ to predict dependent variable(y) of the train dataset is shown below,

$$E(f) = \sum_{i=1}^{N} L(y_i, f(x_i)). \quad (2)$$

The goal is to minimize $E(f)$ with respect to f.

3.2. Cross-Validation and Randomized Search CV

To make sure the algorithms have the optimal performance on the working dataset, Cross-validation is used, which is a resampling method that uses different parts of the data to test and train the model in different iterations. The purpose of performing CV is to assess the model's ability to predict new data that was not used in the training.

Randomized Search CV is implemented throughout the analysis, Randomized Search CV randomizes search on hyperparameters, it is useful when there are many parameters to try, and the training time is relatively long [13, 14].

In terms of spotting potential problems for Cross-Validations such as over-fitting and selection bias, performance metrics will be used in the next step to compare the outcomes of each algorithm. This paper uses Root Mean Square Error, Mean Absolute Percentage Error and R-squared to summarize the errors.

Also note that cross-validation can be problematic for time-series models, but since the time factor is not relevant in this analysis, this can be ignored.

3.3. Performance metrics

To evaluate the result of each algorithm, Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE) and R-squared (R2) are used.

3.4. RMSE

RMSE is requently used to measure the difference between values preducred by a model or estimated values and the observed values, which represents the square root of the sample moment of the difference between the predicted value and the observed value or the quadratic mean of these differences. RMSE is a measure of accuracy, comparing forecasting errors of different models for a particular dataset [15]. The formula of RMSE is as follows,

$$RMSE = \sqrt{\frac{1}{n}\sum_{t=1}^{n}(y_t - \hat{y}_t)^2} \quad (3)$$

3.4.1. MAPE

Mean absolute percentage error (MAPE) is a measure of prediction accuracy of a forecasting method. MAPE is

usually used as a loss function for regression problems and in model evaluation. Since it is intuitive in terms of relative error, in the scope of consistency, MAPE is feasible as a loss function for regression because the existence of an optimal model and the consistency of the empirical risk minimization can be proved [16]. The formula of MAPE is shown below,

$$MAPE = \frac{1}{n}\sum_{t=1}^{n}\left|\frac{y_t - \hat{y}_t}{y_t}\right| \quad (4)$$

3.4.2. R-squared

The coefficient of determination, also known as R-squared, shows how the difference in one variable is explained by the difference in a second variable in predicting the outcome of a given event, or how strong is the linear relationship between the two variables [17]. R-squared gives us information about how well the fit of a model is, therefore comparing R-squared value between algorithms allows us to see how each algorithm fits the data.

The formula for R-squared is shown below,

$$R^2 = 1 - \frac{\sum_i e_i^2}{\sum_i (y_i - \bar{y})^2} \quad (5)$$

4. EXPERIMENT RESULTS

4.1. Data processing and Exploratory Data Analysis

To facilitate processing of data sets, pre-processing was done before any machine learning can be applied. Pre-processing is important because it prepares the data in a meaningful and valuable way for the following detailed analysis, and the quality of pre-processing decides the outcomes since algorithms depend heavily on the proper data it needed to learn [17].

Firstly, maximum and minimum cost prices are replaced by a new variable average cost price, considering vehicles tend to have outliers that have a relatively large distance to mean, variables such as maximum and minimum cost price are not very helpful in terms of analysis. For the same reason brands that contain less than 10 values were dropped.

Standardization is the process of converting data into a uniform format that allows us to manipulate the data easily. Since some of the variables in our dataset had high variances and might be weighted more heavily, standardization was used to standardize each variable to a comparable scale with a mean of 0 and a standard deviation of 1 [17, 18]. After standardization, we create dummy variables for 3 categorical variables in the dataset, seller type, fuel type and transmission type. This transformation simplifies the analysis by allowing us to use a single regression equation to represent multiple groups rather than having to write a separate equation for each subgroup.

TABLE IV. PERFORMANCE METRICS OF EACH MODEL

Model	Train			Cross-Validation (10 folds)			Test		
	RMSE	MAPE	R2	RMSE	MAPE	R2	RMSE	MAPE	R2
LR	454.65	0.48	0.68	453.11	0.48	0.68			
DTR	12.69	0.0047	0.9998	242.69	0.16	0.91			
RFR	67.62	0.05	0.99	180.57	0.12	0.95	192.10	0.12	0.94
GBR	156.44	0.15	0.96	188.02	0.15	0.94	172.51	0.12	0.95

Notes: LR represents Linear Regression, DTR is Decision Tree Regression, RFR represents Random Forest Regression, GBR represents Gradient Boosting Regression.

4.2. Hyperparameter optimization

To get the model that can fit the dataset and predict optimally, we use Randomized Search CV with 5 folds validation as a hyperparameter optimization or tuning method to find an optimal tuple of hyperparameters for the model. This method may help improve the overfitting problem.

Because Linear Regression Model and Decision Tree Regression Model performed badly during our analysis, we only applied Hyperparameter optimization on Random Forest Regression Model and Gradient Boosting Regression Model.

For Random Forest Regression Model, we create two sets of adjustable parameters, the number of trees in the forest (n_estimators) is set as [3, 10, 30, 60] for both sets and the number of features to consider when looking for the best split (max_features) is set as [2, 4, 6, 8] for both sets. The only difference between the two sets is bootstrap is set as True in set 1 by default but False in set 2.

For Gradient Boosting Regression Model, we only set one tuple with learning rate = [0.01, 0.05, 0.1], the number of boosting stages (n_estimators) = [100, 150, 200, 250, 300], maximum depth of the individual regression estimators = [2, 5, 10, 15, 20, 30], minimum number of samples required to be at a leaf node (min_samples_leaf) = [1,5,10,20,30,50], and minimum number of samples required to split an internal node (min_samples_split) = [2,5,10,20,40,50].

4.3. Prediction power comparison

From simple linear models to complex machine learning models, Linear, Decision Tree, Random Forest and Gradient Boosting regression models are selected to compare. To avoid overfitting problems and ensure the result is not biased, the commonly used number of folds is set at 10 for Cross-Validation.

In Table 4, Linear Regression model performed worst with the highest RMSE and MAPE, and the smallest R2. Although Train RMSE, MAPE and R2 of Decision Tree Regression all show excellent performance and low error, its Cross Validation RMSE, MAPE and R2 are all the second worst. The huge difference in its performance in Train dataset and Cross-Validation indicates that the Decision Tree Regression model is suffered from an overfitting problem. Random Forest Regression model has the same overfitting problem as the Decision Tree Regression Model, but it performed best in Cross-Validation. The last Gradient Boosting Regression Model has no overfitting problem and shows great performance.

Because Linear Regression, and Decision Tree Regression models performed badly, only Random Forest and Gradient Boosting Regression models are chosen to optimize by using Randomized Search CV described in Section 3.2. After hyperparameter optimization, we achieve two final models, which are used to fit the test dataset and compare the strength of predictions. The result shows that the Gradient Boosting Regression model has better performance in the test dataset than that of Random Forest Regression model.

To conclude, although Gradient Boosting Regression Model did not perform best in Train dataset and after cross-validation respectively, it has very stable RMSE, MAPE and R2 which indicate that there is no overfitting problem. Also, its prediction performance in test dataset remains stable and shows great prediction power. Based on the analysis above, we choose Gradient Boosting Regression Model as our final optimal model.

4.4. Feature importance in the Final Model

Feature importance indicates how useful the feature or independent variables are at predicting the target variable. This helps to know which variables have a big impact on used car prices in our model. Figure 8 shows the top 6 features that contribute to the prediction of used car price in the Final Gradient Boosting Regression Model. max_power is the most important feature in the final optimal model followed by vehicle_age, avg_cost_price and km_driven.

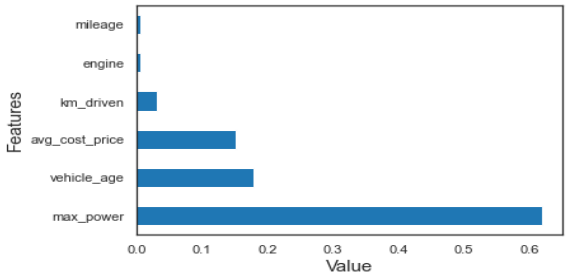

Figure 8. Feature Importance

This suggests that a single variable, max_power, can largely determine used-car prices, considering other features may lead to an accurate result.

5. CONCLUSION

In this paper, to build a model that can predict the used car prices in India, and to help solve the information asymmetry problem in the traditional 'lemon' used car market, exploratory data analysis and data preprocessing are carried out to understand the main characteristics of the dataset and to make data more compatible to the machine learning algorithms. Then, four machine learning algorithms, including Linear Regression, Decision Tree Regression, Random Forest Regression and Gradient Boosting Regression, are utilized to build models. After fitting the train and test dataset, selecting the model by Cross-validation, optimizing hyperparameters by Randomized Search CV, Gradient Boosting Regression model is selected as the optimal model because it shows great prediction power and has no overfitting problem.

Based on the performance of our final optimal model, our final gradient boosting model can be directly used to predict used car prices in India market at the current point. Also, max_power has the highest feature importance value, it can already give a rough prediction of used car price but inputting other variables will lead to a more accurate predicted price. Then, because our final optimal model is backed by more than 15,000 data, 10 different factors, and analyzed through machine learning algorithms, buyers can get much more information about used car prices in India, and sellers are hard to hide information in face of our relatively comprehensive model. Thus, Information asymmetry and lemon market problems can be solved to a certain extent by using and continuously updating our final optimal model.

However, there are still some limitations in this paper, we used a dataset containing more than 15,000 used cars, but it is still not enough. In the preprocessing step, we

drop several brands of cars because we have little information about them, if we can a bigger dataset that contains more used cars, our model will perform better and be more comprehensive.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

Shiyu Gu: Methodology, Modelling, Python Code, Investigation, Writing.

Ximing Deng: Methodology, Investigation , Writing.

Xinru Chen: Data Analysis, Investigation, Writing.

Lei Huang: Data Collection, Analysis, Writing, Supervision.

REFERENCES

[1] Hedge Company. (2021) How many cars are there in the world in 2021? stats by country [Online]. Available: https://hedgescompany.com/blog/2021/06/how-many-cars-are-there-in-the-world/

[2] Mordor Intelligence. (2021) Used car market - growth, trends, COVID-19 impact, and forecasts (2021-2026) [Online]. *Mordorintelligence.com*. Available: https://www.mordorintelligence.com/industry-reports/global-used-car-market-growth-trends-and-forecast-2019-2024

[3] R. Hensley, I. Maurer, and A. Padhi. (2021) How the automotive industry is accelerating out of the turn [Online]. Available: https://www.mckinsey.com/industries/automotive-and-assembly/our-insights/how-the-automotive-industry-is-accelerating-out-of-the-turn

[4] G. A. Akerlof. "The Market for Lemons: Quality Uncertainty and the Market Mechanism," *The Quarterly Journal of Economics*. Vol. 84, no. 3, pp.488-500, 1970.

[5] W. Emons, and G. Sheldon. "The market for used cars: new evidence of the lemons phenomenon," *Applied Economics.*, vol. 41, no. 22, pp. 2867-2885, Oct. 2009.

[6] L. Zavolokina, G. Miscione, and G. Schwabe, "Buyers of 'lemons': How can a blockchain platform address buyers' needs in the market for 'lemons'?," *Electron Markets*, vol. 30, no. 2, pp. 227–239, 2019.

[7] X. Yan, and X. G. Su, *Linear Regression Analysis: Theory and Computing*, World Scientific, 2009, pp. 1-2.

[8] X. Wu, V. Kumar, Q. J. Ross, et al. "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp.1-37, 2008.

[9] Y. Song, and Y. Lu, "Decision tree methods: applications for classification and prediction," *Shanghai Arch Psychiatry*., vol. 27, no. 5, pp. 130-135, 2015.

[10] A. Liaw, and M. Wiener, "Classification and Regression by RandomForest," R News, vol. 2, no. 3, pp. 18-22, 2002.

[11] S. M. Piryonesi, and E. E. Tamer, "Data Analytics in Asset Management: Cost-Effective Prediction of the Pavement Condition Index," *Journal of infrastructure systems*, vol. 26, no. 1, pp.4019036, 2020.

[12] T. Hastie, R. Tibshirani, and J. H. Friedman, *The Elements of Statistical Learning*, 2nd ed. New York: Springer, 2009, pp. 337–384.

[13] R. J. Hyndman, and A. B. Koehler. "Another look at measures of forecast accuracy," *International Journal of Forecasting.*, vol. 22, no. 4, pp. 679-688, 2006.

[14] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Transactions on Neural Networks*, vol. 4, no. 4, pp. 570-578, July 1993.

[15] J. S. Armstrong, and F. Collopy, "Error Measures For Generalizing About Forecasting Methods: Empirical Comparisons," *International Journal of Forecasting*, vol. 8, no. 1, pp. 69–80, 1992.

[16] B. de Myttenaere, B. Golden, B. Le Grand, and F. Rossi, "Mean absolute percentage error for regression models," *Neurocomputing (Amsterdam)*, vol. 192, pp. 38-48, 2016.

[17] Allen, and M. David, "The Relationship between Variable Selection and Data Agumentation and a Method for Prediction". *Technometrics*, vol. 16, no. 1, pp. 125–127, 1974.

[18] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," Journal of the Royal Statistical Society, Series B (Methodological), vol 35, no. 2, pp.111–147, 1974.

***S. Gu** was born in Zhejiang Province of China in 1999. Gu is currently pursuing a Bachelor of Commerce degree at The University of Sydney Business School in Australia, majoring in Business Analytics and Finance.

He previously worked as an INTERN at Digital Solutions Department in PwC Shanghai for three months and as a PROJECT ASSISTANT at Blue Emerging Markets Division in Bosch Power Tools for four months. He is a Dalyell Scholar at The University of Sydney and focuses on data science and operations research.

X. Deng was born in Sichuan Province of China in 2000. Deng is currently pursuing a Bachelor of Accountancy and Bachelor of Information System degree at Gies College of Business, University of Illinois at Urbana-Champaign, United States. Her study focused on the technical knowledge in accounting, assurance, and data analytics in business.

She previously worked as an INTERN at E-Commerce Department at Beijing Inplayable Inc. for three months and as an ACCOUNTANCY AND FINANCIAL ANALYST at IdeaMill 19 Inc. in United States for one year. She worked as a student consultant at Illinois Business Consulting during her sophomore year.

X. Chen was born in Guangdong Province of China in 2000. Chen is now pursuing her bachelor degree at University of Washington in Seattle, United States. She is majored in Applied and Computational Mathematical Science, with a focus on Scientific Computing & Numerical Algorithms.

In summer 2021, she worked as an INTERN at Huawei Technologies Co., Ltd as a Software Tester for two months. She also worked as a DATA ANALYST in the Internship in Tencent Technology (ShenZhen) Company Limited in summer 2019 for 3 months.

L. Huang was born in Zhejiang Province of China in 1999, Huang is currently pursuing a Bachelor of Applied Physics degree at Physics Department of University of Washington. His study focused on data analysis in science experiments.

Huang has previously worked as an INTERN in Ipsos as a Market Researcher and Data Analyst for over two months. He has also worked as INTERN at Zhongtai Securities Co. as a Data Analyst.