Prediction And Investigation for U.S. Used Car Prices and Factors Related to Price

Muyuan Li *

Department of Computer Science, University of Arizona, Tucson, United States

* Corresponding Author Email: muyuanl@arizona.edu

Abstract. In contemporary society, as transportation infrastructure continues to advance, a growing number of families opt for car ownership as a means of commuting. However, constrained by budget considerations, many households opt for pre-owned vehicles. This underscores the significance of comprehending the dynamics of the used car market and delving into the factors that influence the pricing of these previously owned automobiles. The author found a data set of used cars in United States on Kaggle and used it for analysis. This dataset contains data about 762,091 used cars scraped from cars.com. The data was collected in Apr, 2023. The research question is how to predict the used car prices and what factors are related to the used car prices. The author uses Linear Regression and Random Forest Regression to predict used car prices. The author found the factors related to the price of used cars as well. The study found that the Random Forest Regression method is more accurate in predicting used car prices. A total of 17 used car factors correlated with used car price. Such as mileage, car age, mpg and so on.

Keywords: Used car; Linear Regression; Random Forest Regression; Factors; Price.

1. Introduction

The paper begins with an introduction that provides an overview of the problem at hand - predicting used car prices. The author introduces two methods used for prediction - Linear Regression [1] and Random Forest Regression [2]. These are popular machine learning algorithms that have been proven to be effective in various prediction tasks.

In Section 2, the author delves into the methodology used in the study. This section begins with an explanation of the process of Exploratory Data Analysis (EDA) [3]. EDA is a crucial step in any data analysis task as it allows for a better understanding of the data at hand. The author explains how this process was carried out and provides insights into the data.

This section also discusses how the two models were trained. Training a model is a complex process that involves feeding the model with data and allowing it to learn from it. The author explains this process in detail.

The correlation heat map of factors related to used car prices is also presented in this section. A correlation heat map is a graphical representation of the correlation matrix, which shows the correlation coefficients between various variables. In this case, it shows how different factors are related to used car prices. This visual representation makes it easier to understand the relationships between different variables and their impact on used car prices. Section 3 presents the results of the study. The accuracy of the two models is discussed here. Accuracy is a key metric in evaluating the performance of a prediction model. The higher the accuracy, the better the model is at making predictions. The author compares the accuracy of Linear Regression and Random Forest Regression, providing a clear picture of which model performs better. The scatter plot of their predicted results provides a visual comparison of how well each model performed. Scatter plots are great for visualizing relationships between two variables, in this case, the actual vs predicted prices. Each point on the plot represents a car, with its actual price on one axis and its predicted price on the other. The author also ranks factors related to used car prices in order of importance. This ranking gives us some insight into which factors have a greater impact on used car price. Section 4 contains all literature cited in this paper, providing readers with resources for further reading and understanding. Finally,

Section 5 provides a summary and conclusion of what has been done in this study. It revisits key points made throughout the paper and suggests areas for further exploration or future work.

Overall, this paper provides an in-depth analysis of predicting used car prices using Linear Regression and Random Forest Regression. It offers insights into how different factors influence used car prices and compares the effectiveness of two popular machine learning models in making these predictions.

2. Methodology

Figure 1 shows the mind map of the methodology [4] the author used for the paper.

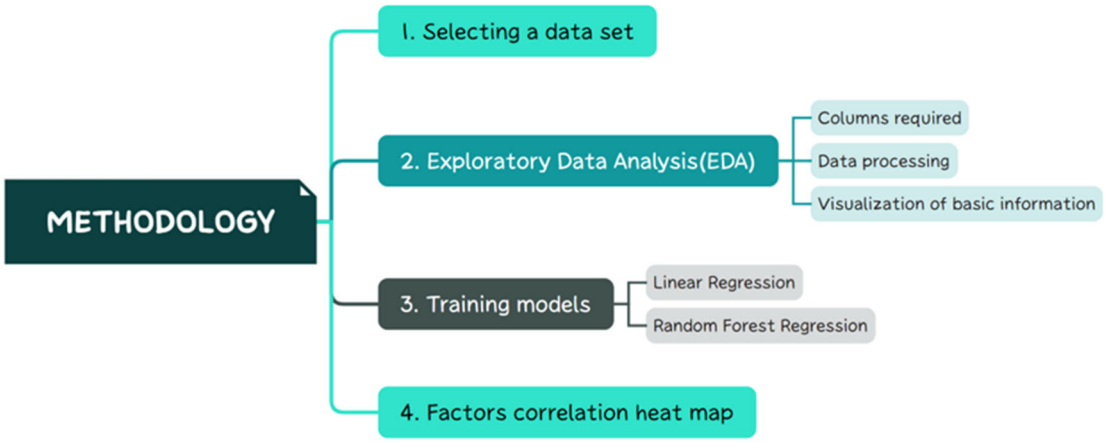


Figure 1. Mind map of the paper (Photo/Picture credit: Original)

2.1. Select a data set

The author chose a data set on Kaggle [5]. The data source for this data set is cars.com. There are 20 columns in total. Such as manufacturer, model, mileage, engine, transmission and so on.

2.2. Exploratory Data Analysis(EDA)

2.2.1 Columns required

After observing these data, the author dropped 2 columns: seller_name and price_drop. The rest of these columns are related to predicting used car prices: manufacturer, model, year, mileage, engine, transmission, drivetrain, fuel_type, mpg, exterior_color, interior_color, accidents_or_damage, one_owner, personal_use_only, seller_rating, driver_rating, driver_reviews_num and price. There are 18 columns in total.

2.2.2 Data processing

The author uses dropna() method to drop rows that are missing data. And the author uses drop_duplicates() method to drop duplicate rows.

The author has manipulated the 'year' column to change the year of manufacture of the used car to the age of the used car. The author observed that the 'price' and 'mileage' columns in the data set had many unusual values. For example, many prices have the value of 0. Therefore, the author specify a range of values for price and mileage. The author set the upper limit of the price at 95% of the size of the value in its column and the lower limit at 5% of the size of the value in its column. The author did the same for mileage. By using the 'shape' method, there are 336,497 rows of data in the current data set.

2.2.3 Visualization of basic information

Figure 2 shows the price box plot.

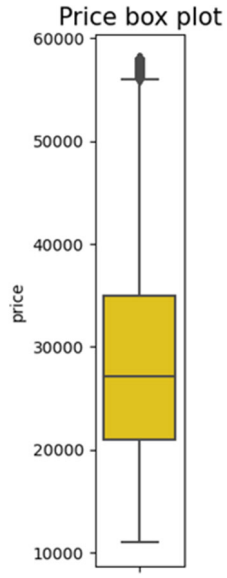


Figure 2. Price box plot (Photo/Picture credit: Original)

From bottom to the top:
Lower limit: $10,989.0
Lower quartile(Q1): $20,978.0
Median(Q2): $27,099.0
Upper quartile(Q3): $35,000.0
IQR: $14,022.0
Upper limit: $57,994.0
Figure 3 shows the mileage box plot.

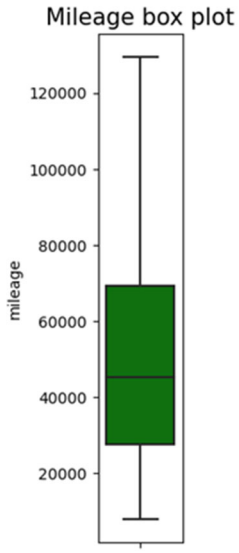


Figure 3. Mileage box plot (Photo/Picture credit: Original)

From bottom to the top:
Lower limit: 7,897.0(mi)
Lower quartile(Q1): 27,705.0(mi)
Median(Q2): 45,130.0(mi)
Upper quartile(Q3): 69,426.0(mi)
IQR: 41,721.0(mi)
Upper limit: 129,428.0(mi)
Figure 4 shows the distribution of car age:

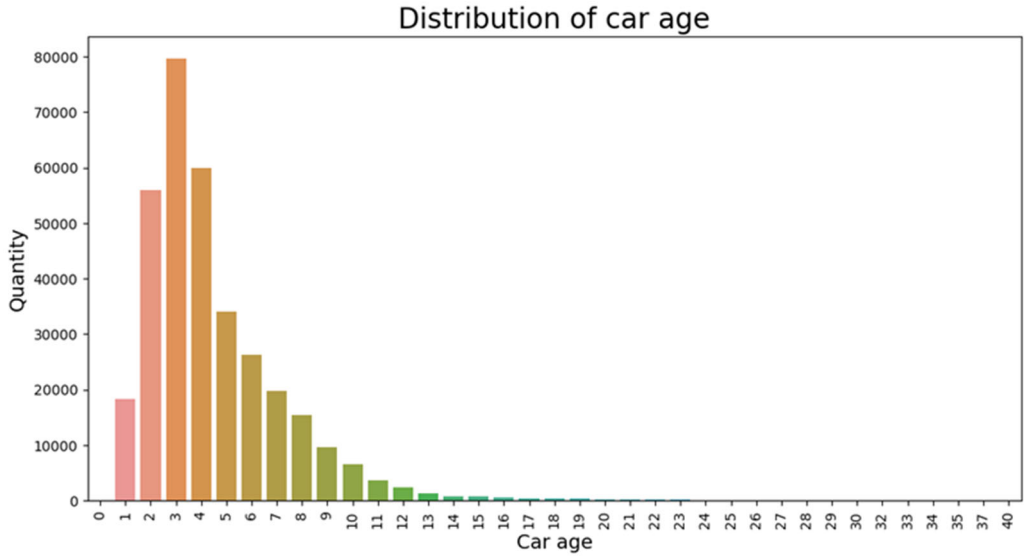


Figure 4. Distribution of car age (Photo/Picture credit: Original)

It is known from observation that most used cars are around 3 years old.

2.3. Training models: Linear Regression and Random Forest Regression

The linear regression and random forest regression model were employed. Even though their performance may not be as good as the current popular neural networks widely used in many studies [6-9], they have better interpretability and are easy to analyze and predict the results. In the author's code, `X` and `y` are two variables that represent the independent and dependent variables, respectively. The column corresponding to 'y' is 'price'. In order to save time and improve accuracy, the author drops the columns “manufacturer”, “interior_color” and “exterior_color” in addition to price.

The author used 'get_dummies()' function to convert categorical variables into dummy variables (also known as indicator variables) for use in the modeling process. The program used the factors in `X` to predict used car prices `y`.

The author used the 'train_test_split()' method to split the training set and test set. Where the size of the test set is 10% of the data set.

Train the Linear Regression model using the 'fit()' method. Model training takes about 10 minutes or more.

The author did the same thing for Random Forest Regression, but also dropped the “model” and “engine” columns. Since model training is too slow, the author used “n_jobs=-1” let the model use all available processors for training, which can speed up the training process. The author changed the test set from 10% to 20% of the data set as well. Currently the author can train in less than 10 minutes.

2.4. Factors correlation heat map

Figure 5 shows the correlation heat map [10]. The darker the color, the stronger the relationship between the 2 factors.

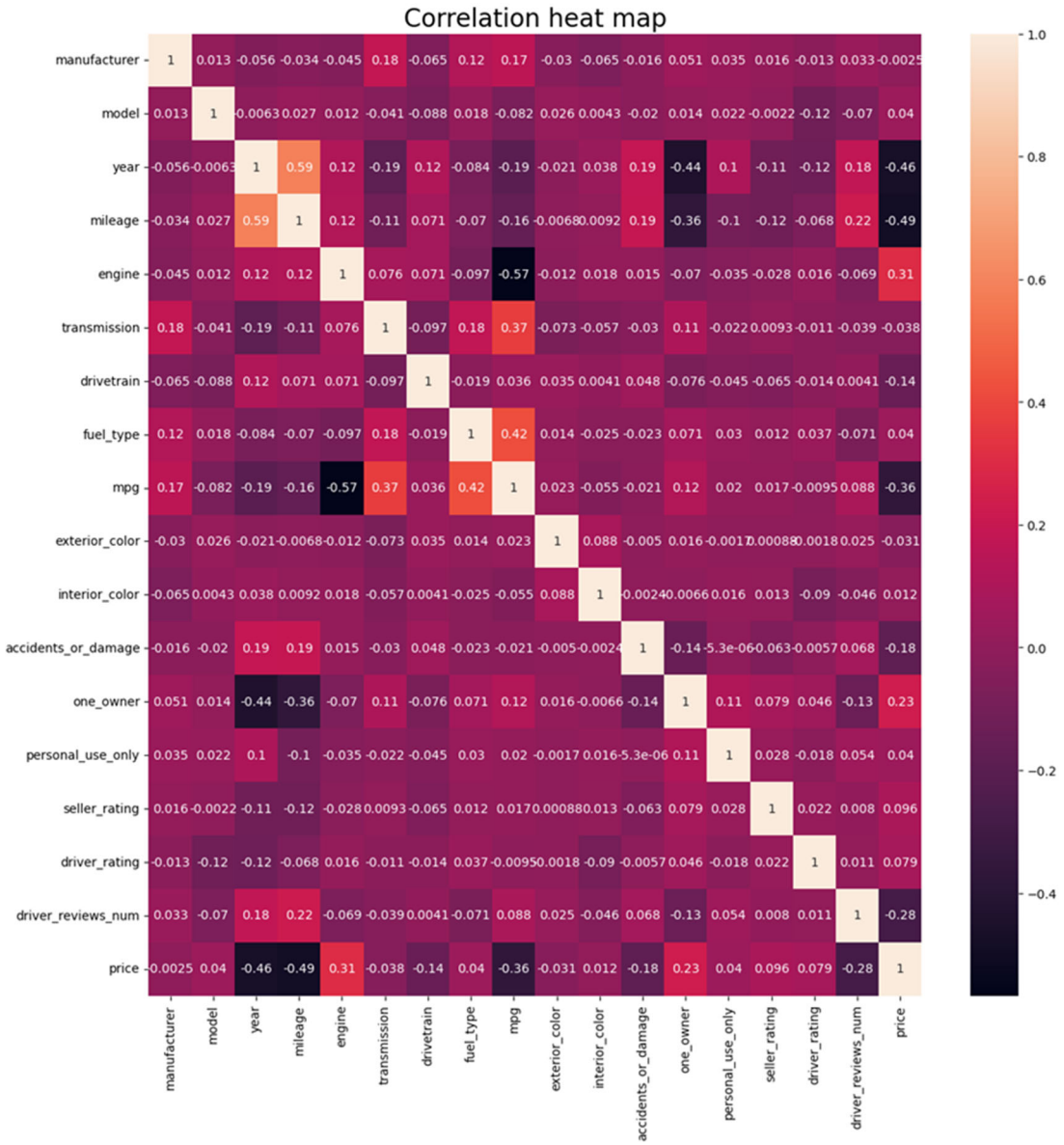


Figure 5. Correlation heat map (Photo/Picture credit : Original)

3. Results and Discussion

3.1. Comparison of 2 models

The author visualized the 2 models using scatter plots. In Figure 6 and Figure 7, the blue and red points in the graph indicate the intersection of the predicted and true values of used car prices. If the predicted price differs from the actual price by less than $1,000, the points are blue.

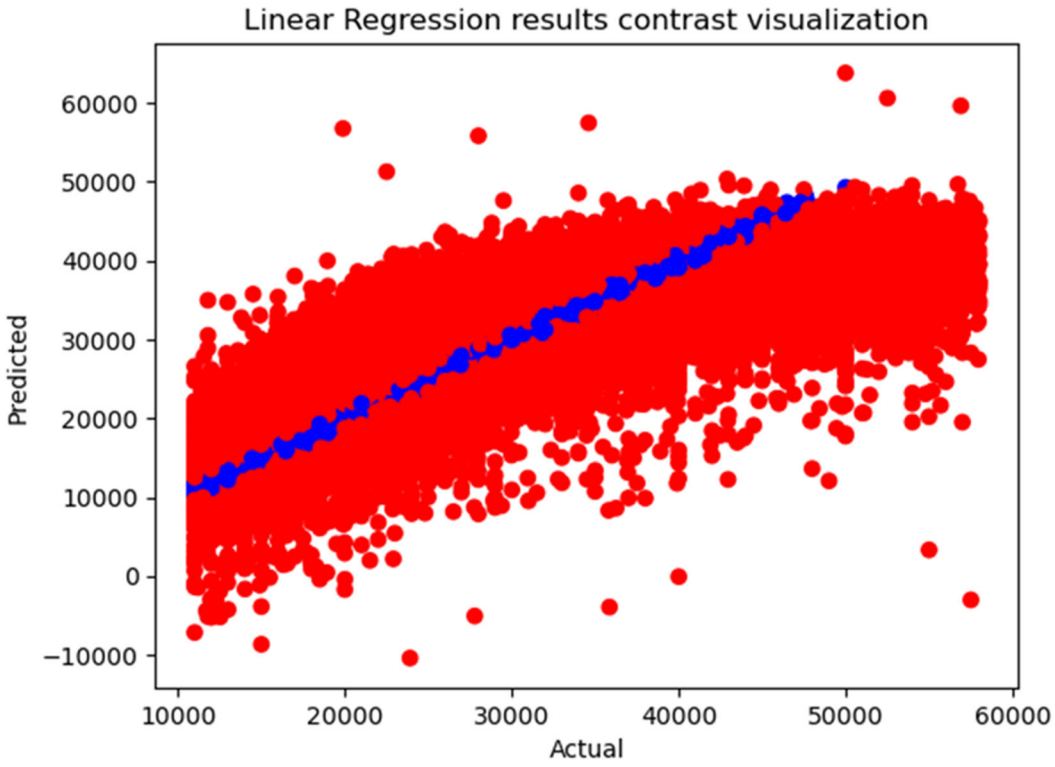


Figure 6. Linear Regression results contrast visualization

The accuracy(R-Squared) of the Linear Regression model in this run is about 57.57%. There are 29,329 red points and 4,321 blue points. As can be seen from the Figure 6, the Linear Regression model sometimes predicts a negative price. Some outliers vary widely.

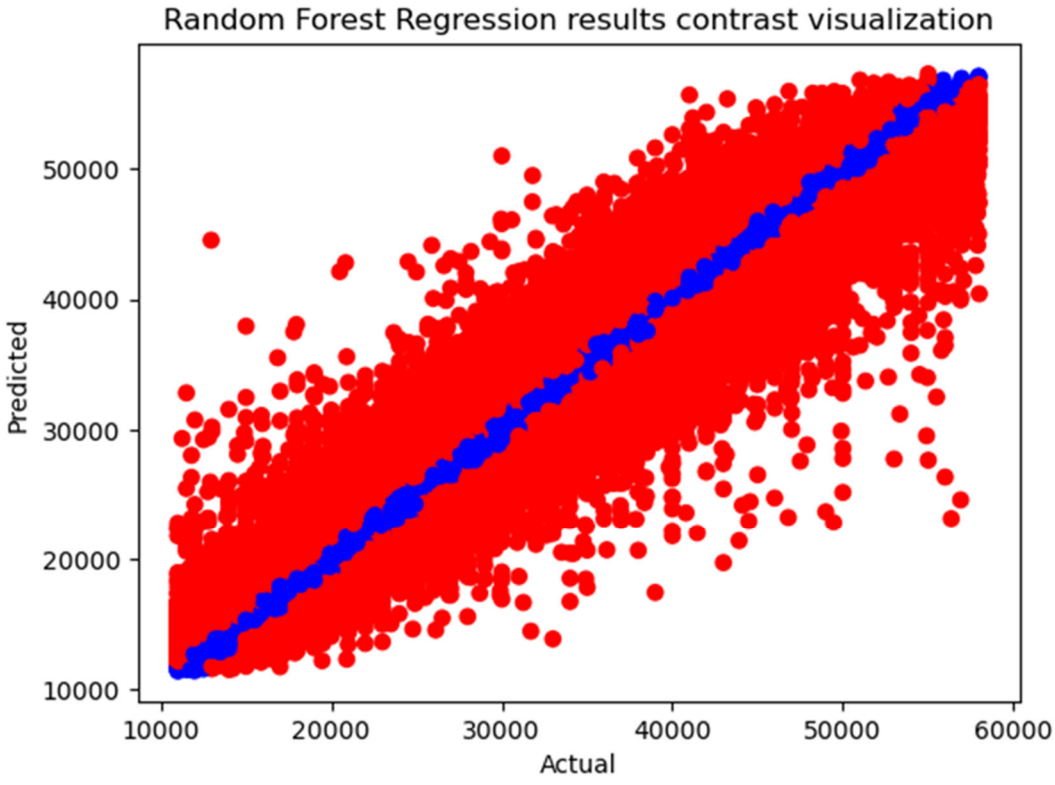


Figure 7. Random Forest Regression results contrast visualization

The accuracy(R-Squared) of the Random Forest Regression model in this run is about 90.97%. There are 45,208 red points and 22,092 blue points. As can be seen from the Figure 7, the Random Forest Regression points is more dense than Linear Regression. No predicted value is negative.

The comparison of the 2 models is shown in Table 1:

Table 1. The performacne of models.

Comparison item	Linear Regression	Random Forest Regression
Red points	29,329	45,208
Blue points	4,321	22,092
Accuracy	57.57%	90.97%

3.2. Factors related to used car price

The author took absolute values for the related parameters of these factors. Factors of used car related to price are listed below in order of correlation from largest to smallest:

1. mileage
2. year
3. mpg
4. engine
5. driver reviews num
6. one owner
7. accidents or damage
8. drivetrain
9. seller rating
10. driver rating
11. model
12. fuel type
13. personal use only
14. transmission
15. exterior color
16. interior color
17. manufacturer

4. Conclusion

The author predicted the price of used cars using 2 models: Linear Regression and Random Forest Regression. The author identified the factors associated with the price of used cars. Through the study, this paper found that the Random Forest Regression model is much more accurate than Linear Regression. The Factors that are most associated with used car prices are mileage, car age and oil consumption. The main contribution of this paper is to compare Linear Regression and Random Forest Regression and rank the factors associated with used car prices. This paper can serve as a reference for people who want to buy used car and other related researchers. The shortcutting of this paper is that the author only compared the prediction results of 2 models. The author believes that the accuracy of other models in predicting prices can be compared in future studies.

References

[1] Hope T M, Linear regression. In Machine Learning (pp. 67-81). Academic Press, 2020.

[2] Segal M R, Machine learning benchmarks and random forest regression, 2004.

[3] Behrens J T, and Yu C H. Exploratory data analysis. Handbook of psychology, 2, 33-64, 2003.

[4] Sileyew K J. Research design and methodology. Cyberspace, 2019, 1-12.

[5] Andrei N, Used Cars Dataset. Kaggle. Available at: https://www.kaggle.com/datasets/andreinovikov/used-cars-dataset, 2023.

[6] Brause R W. Medical analysis and diagnosis by neural networks. Medical Data Analysis: Second International Symposium, ISMDA 2001 Madrid, Spain, October 8–9, 2001 Proceedings 2. Springer Berlin Heidelberg, 2001: 1-13.

[7] Qiu Y, et al. Pose-guided matching based on deep learning for assessing quality of action on rehabilitation training. Biomedical Signal Processing and Control, 72, 103323, 2022.

[8] Brause R W. Medical analysis and diagnosis by neural networks. Medical Data Analysis: Second International Symposium, ISMDA 2001 Madrid, Spain, October 8–9, 2001 Proceedings 2. Springer Berlin Heidelberg, 2001: 1-13.

[9] Kaastra I Boyd M. Designing a neural network for forecasting financial and economic time series. Neurocomputing, 1996, 10(3): 215-236.

[10] DeBoer M. Understanding the heat map. Cartographic perspectives, 2015 (80): 39-43.