**CAR PRICE PREDICTION USING MACHINE LEARNING TECHNIQUES**

Sanjana Arkatala

*Computer science*

*University of Central Missouri*

Lees Summit, Missouri

sxa71570@ucmo.edu

Chandra Mouli Krishna Vamsi Mendru

*Computer science*

*University of Central Missouri*

Lees Summit, Missouri

Cxm10690@ucmo.edu

***Abstract* —** Using machine learning to determine car prices is analogous to how expert systems acquire new information. In order to learn about cars, the most common method now is to read recommendations and posts made on online market websites. After locating the data, it can be classified as either "structured" (already arranged) or "unstructured" (needing to be reorganized) to facilitate analysis in light of our existing expertise. Meaning extraction, data inference, and qualitative data rules are just a few of the methods we'll cover here. The major purpose of this study is to analyze a wide range of data pertaining to automobiles in order to develop a mechanized method for estimating future automobile pricing. The auto industry is becoming increasingly global and competitive annually. As a result, in today's cutthroat auto industry, it is essential that fair prices be established for both consumers and producers. There is a lack of clarity regarding the appropriate purchase or selling price of automobiles, both among consumers and among manufacturers. As a result, both consumers and businesses look to auto dealers, periodicals, and online resources for guidance. This data, however, is time-consuming to compile and may cause market confusion. There are a lot of moving parts in the car market, thus trying to predict their costs is seen as difficult. Car prices are affected not only by internal criteria like make, model, year, engine size, gas mileage, and so on, but also by a wide range of governmental levies and road conditions (for used car sales). Therefore, the buying and selling of automobiles is a huge and diverse economic activity. E-commerce has made it simpler for buyers and sellers to transact business, and this includes the buying and selling of automobiles.

Divyanjali Gundala

*Computer science*

*University of Central Missouri*

Lees Summit, Missouri

dxg36740@ucmo.edu

Tasneem Shaik

*Computer science*

*University of Central Missouri*

Lees Summit, Missouri

Txs10260@ucmo.edu

**Keywords: Car Price prediction, Regression, analysis, random forest, linear regressor**

# Introduction

We conducted a comparative study on performance of regression based on supervised machine learning models. Each model is trained using data of used car market collected from German e-commerce website. Considering the demand for private cars all around the world the demand in the second-hand car market has been rising and creating a chance in business for both buyer and seller. In several countries, buying a used car is the best choice for customers because its price is reasonable and affordable to the buyer. After a few years of using them, it may get a profit from reselling again. However, various factors influence the price of a used car such as how old of those vehicles and the condition in current scenario of them. Normally, the price of used cars in the market is not constant. Thus, car price evaluation model is required for helping in trading. In this paper, we conducted a comparative study using multiple linear regression, random forest regression and gradient boosted regression trees to build a price model of a used car. Each algorithm used data scraped from e-commerce websites. The primary objective of this project is to find the best predictive model for predicting used car prices. In order to learn about cars, the most common method currently is to read suggestions and posts made on online market websites. Once the data is located, it is categorized as "structured" if it is already organized or "unstructured" if it has to be rearranged to be analyzed using our current knowledge. Several techniques, including meaning extraction, data inference, and qualitative data rules, will be discussed. The primary goal of this research is to create an automated system for predicting future automotive price by analyzing a large amount of data relating to cars. Competition and globalization in the car business are rising every year. Given the current level of competition in the car business, it is crucial that reasonable pricing be set for both buyers and sellers. Neither car buyers nor sellers have a firm grasp on what constitutes a fair price for a vehicle. Therefore, both individuals and companies seek advice from car lots, magazines, and the Internet. But gathering this information can be tedious, and it might potentially lead to market misunderstanding. Predicting automobile prices is considered challenging due to the market's many variables. Internal factors such as make, model, year, engine size, gas mileage, and so on are certainly influential, but external factors such as government taxes and road conditions also play a role in determining the final sticker price of a car (for used car sales).

# Motivation

Expert system knowledge acquisition and machine learning are intimately connected in the context of setting automotive rates. More often than not, people have to go through the trouble of putting suggestions and ads for cars they want to buy or sell on internet marketplaces just to get the information they need. Once located, the data can be classified as structured and straightforward to evaluate or unstructured and puzzling. The auto business is becoming increasingly global and competitive annually. As a result, in today's cutthroat auto industry, it is essential that fair prices be established for both consumers and producers. There is a lack of clarity on the appropriate purchase or selling price of automobiles, both among consumers and among manufacturers. As a result, both consumers and businesses go to auto dealers, periodicals, and online resources for guidance. This data, however, is time-consuming to compile and may cause market confusion. As the quality of mobile Internet continues to rise, conventional offline models for buying and selling used cars are becoming more out of date. In response, online marketplaces for buying and selling pre-owned vehicles have evolved. The foundation of the used automobile market is the process of determining a fair price for the vehicle, which should accurately reflect the market's objective, fair conditions. The linear correlation between vehicle parameters, vehicle conditions, and transaction factors and used car price was extensively explored, grey relational analysis was applied to filter the feature variables of factors affecting used car price, and the traditional BP neural network was optimized by combining the particle swarm optimization and a support vector machine to improve the accuracy of used car price forecasts.

# Main Contributions & Objectives

1. Detect and remove outliers in numerical variables
2. Drop and fill missing values
3. Feature Engineering
4. Multi Linear Regression
5. Lasso Regression
6. Support Vector Regression
7. Random Forest Regression

# Related Work

There have been a number of prior publications on the topic of estimating the value of old automobiles. Pudaruth [1] multivariate logistic regression, k-nearest neighbors, naive Bayes, and decision trees were used to forecast the cost of a used automobile in Mauritius. However, their results were not useful for forecasting because they saw fewer cars. In his article, Pudaruth drew the conclusion that the decision tree and naive Bayes cannot be used for a continuous value variable.

Noor and Jan [2] used multiple linear regression to predict vehicle car price. They performed variable selection technique to find the most influencing variables then eliminate the rest. The data contain only selected variable that used to form the linear regression model. The result was impressive with R-square = 98%.

Peerun et al. [3] conducted a study to determine how well a neural network can forecast the value of old automobiles. However, especially for more expensive vehicles, the estimated value is far off from the real price. Overall, they found that when comparing support vector machine regression to neural network and linear regression for estimating the value of secondhand automobiles, the former performed somewhat better. Sun et al. [4] suggested implementing the model for estimating the value of secondhand cars purchased online using a BP neural network optimization method. To fine-tune unseen neurons, they presented a novel technique known as the Like Block-Monte Carlo Method (LB-MCM). According to the findings, the optimized model produced more precise results than the unoptimized model. Upon reviewing the relevant literature, we concluded that no prior research have utilized the gradient boosting approach for used vehicle price prediction. Consequently, we choose to construct a model to assess the value of pre-owned automobiles based on a forest of gradient random trees.

1. PROPOSED FRAMEWORK

Specifically, machine learning makes use of regression and classification models. The goal of classification is to assign a yes/no, positive/negative, or other value to the output (or y). One use of sentiment analysis is identifying the review's overall positive or negative tone. However, in regression, we are curious about the value, like the expected price of a house.

The regression model is a reliable tool for projecting the future cost of automobiles. We need to make preparations for a variety of factors, and we also have to cope with a non-normal distribution.

Regression modeling is a tried and true technique for projecting the cost of a car in the future. Many factors require consideration, and we must additionally accommodate for a non-normal distribution.

The output variables in a regression issue are all real numbers. There is a widespread reliance on a linear structure. The following is a common form for the equations of rectilinear regression:

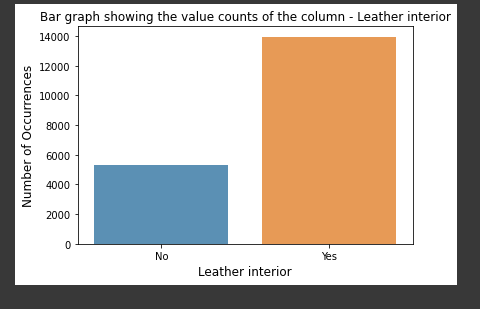
ŷ= w[0] \* x[0] +w[1]\*x[1]+…+w[i]\*x[i]+b

Dual-function versions will make use of the aircraft. If the model has more than two characteristics, then a hyperplane will be used. These are examples of regression algorithms: Regression methods include the well-known ones like linear and logistic regression as well as the less-well-known ones like polynomial regression, extra tree regression, and random forest. Bagging and Ensemble: The ensemble approach is similar, in that several models (commonly called "weak learners") are trained to tackle the same issue and then merged to get superior results; the fundamental premise being that when weak models are appropriately coupled, better results will be achieved.

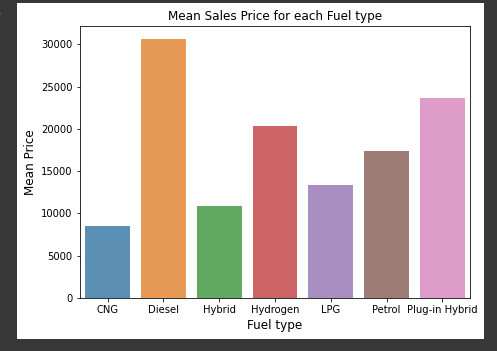
# Data Description

**Analysis on Category:** The category contains 11 types of categories. As per the analysis on occurrence in the dataset we got to know that there is very less data for categories Cabriolet, Coupe, Goods wagon, Limousine, Microbus, Minivan, Pickup, Universal. Further we analyze mean price dependent for every category, and we found that the mean price is almost equal for every category.

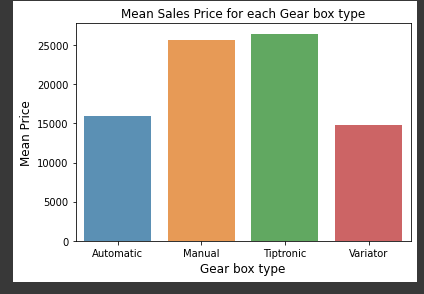
**Analysis on Leather Interior:** We found that there are more cars which have leather interior. We found that the meaning is the same irrespective of its interior.



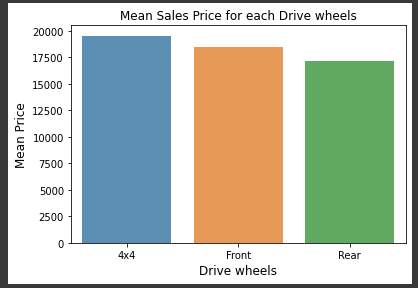
**Analysis on Fuel type:** There are 7 fuel types in our dataset, but we have less data for Hydrogen and Plug-in hybrid. We also found that there is a lot of difference in the Mean of price on the fuel types. Here we found that the fuel type is affecting the price of the car.



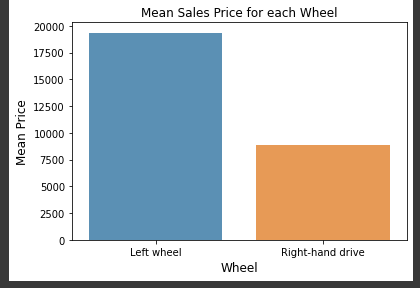
**Analysis on Gear Box:** We found that automatic and variator have similar prices. Similarly, the cars having manual, tiptronic have similar price.



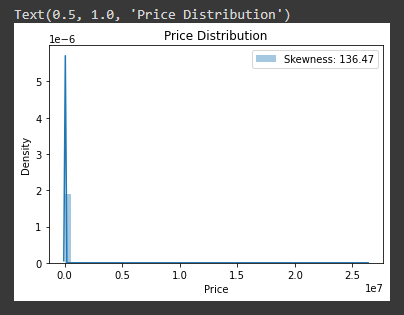
**Analysis on Drive Wheels:** we found that the mean sales price of all the driving wheels is similar.



**Analysis on Wheel:** We found that that cars having a left steering have a higher price than their right counterparts.



**Analysis on Price:**  We found that skewness is very for the price which our target variable.



From the analysis on every column, we found the data needs to be preprocessed before training the model.

1. Detect and remove outliers in numerical variables
2. Drop and fill missing values
3. Feature Engineering
4. Data Transformation
5. Feature Encoding
6. Feature Selection

We used Boxcox Transformation to remove Skewness in numerical columns. Now we look into correlation in the features.

Graphical user interface

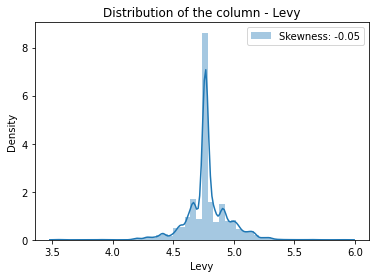
Description automatically generated

# EXPERIMENTATION & ANALYSIS

**Objective 1:**

Extreme numbers that don't fit in with the rest of the data are called outliers. Inaccurate model projections might be a result of outliers, thus it's necessary to deal with them. Tukey's approach is what I'll utilize to get rid of these outliers.

Here, we'll craft a function that iteratively scans a given set of characteristics for anomalies. An outlying data point is one that falls outside the first quartile minus the outlier step or the third quartile plus the outlier step in each iteration. A step that is 1.5 times the interquartile range is considered an anomaly. After the outliers for a given feature have been identified, the feature's index will be added to a list and the procedure will be repeated until all features have been examined. Finally, we'll use the outlier index list to determine how often each index number appears and only return it if it appears more than n times.



**Objective 2:** The dataset has null values only on the column Levy we will be focusing on replacing them in this objective. We filled the column with median.

**Objective 3:**

In this feature engineering we develop a NewCategory such that if the mean price is less than 20000 then it belongs to class 1, else class 2.

Categories with less than 20000 mean price: ['Hatchback', 'Limousine', 'Microbus', 'Sedan'] Categories with more than 20000 mean price: ['Cabriolet', 'Coupe', 'Goods wagon', 'Jeep', 'Minivan', 'Pickup', 'Universal']

**Objective 4:**

**Using multi-linear Regression:**

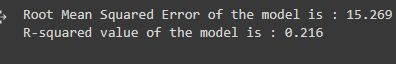
We used linear Regression to train the data. We trained and tested our model using multi linear classification and calculated metric MSE and RMSE metrics.



**Objective 5:**

**Using Lasso Regression:**

To improve the model performance, we used lasso regression. We trained and tested our model using lasso regression and calculated MSE and RMSE metrics. We used 5 sets for cross validation. We got the best parameter are “alpha” = 0.001



**Objective 6:**

**Support vector machine**

To improve the model performance more than lasso regression. We trained and tested our model using support vector machine and calculated MSE and RMSE metrics. We got root mean squared error of the model is 14.105 and R- values is : 0.331

**Objective 7:**

**Random Forest Regressor**

To improve the model performance more than Random Forest Regressor. We trained and tested our model using random forest regressor with different no of trees.

For Trees = 10

Root Mean Squared Error of the model is : 10.62

R-squared value of the model is : 0.621

For trees = 25

Root Mean Squared Error of the model is : 10.324

R-squared value of the model is : 0.642

For trees = 50

Root Mean Squared Error of the model is : 10.149

R-squared value of the model is : 0.654

For trees =100

Root Mean Squared Error of the model is : 10.091

R-squared value of the model is : 0.658

For trees = 1000

Root Mean Squared Error of the model is : 10.04

R-squared value of the model is : 0.661

**Preliminary Results:**

Table

Description automatically generated

# References

[1] B. Kumar and P. Sarkar, “Prediction of future car forms based on historical trends,” Perspectives in Science, vol. 8, pp. 764 – 766, 2016, recent Trends in Engineering and Material Sciences.

[2] S. Lessmann and S. Vo, “Car resale price forecasting: The impact of regression method, private information, and heterogeneity on forecast accuracy,” International Journal of Forecasting, vol. 33, no. 4, pp. 864 – 877, 2017.

[3] J.-D. Wu, C.-C. Hsu, and H.-C. Chen, “An expert system of price forecasting for used cars using adaptive neuro-fuzzy inference,” Expert Systems with Applications, vol. 36, no. 4, pp. 7809 – 7817, 2009.

[4] N. Pal, P. Arora, D. Sundararaman, P. Kohli, and S. Sumanth Palakurthy, “How much is my car worth? A methodology for predicting used cars prices using Random Forest,” arXiv e-prints, p. arXiv:1711.06970, Nov 2017.

[5] E. Gegic et al., “Car price prediction using machine learning techniques,” TEM Journal, vol. 8, no. 1, pp. 113–118, 2019.

[6] K. Noor and S. Jan, “Vehicle price prediction system using machine learning techniques,” International Journal of Computer Applications, vol. 167, pp. 27–31, 06 2017.

[7] Y. Russac, O. Caelen, and L. He-Guelton, “Embeddings of categorical variables for sequential data in fraud context,” pp. 542–552, 2018.

[8] C. Guo and F. Berkhahn, “Entity embeddings of categorical variables,” 2016.

[9] D. S. S. D. S. Yates, Daniel S.; Moore, The Practice of Statistics, 09 2014.

[10] X. Ma, J. Sha, D. Wang, Y. Yu, Q. Yang, and X. Niu, “Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning,” Electronic Commerce Research and Applications, vol. 31, pp. 24 – 39, 2018.

[11] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” 2016

[12] SAS Academy website -“https://www.sas.com/en\_in/insights/analytics/machine-learning.html”

[13] Dr. M. J. Garbade – “Clearing the Confusion: AI vs Machine Learning vs Deep Learning Differences” Available: https://towardsdatascience.com/clearing-the-confusion-ai-vs-machine-learning-vs-deeplearningdifferences-fce69b21d5eb

[14] A. Wilson – “A Brief Introduction to Supervised Learning”

[15] Available:https://towardsdatascience.com/a-brief-introduction-to-supervised-learning54a3e3932590

[16] A. Wilson – “A Brief Introduction to Supervised Learning”

[17] Available:https://towardsdatascience.com/a-brief-introduction-to-supervised-learning54a3e3932590

[18] J. Rocca – “Ensemble methods: bagging, boosting and stacking”

[19] J. Brownlee – “Bagging and Random Forest Ensemble Algorithms for Machine Learning”

[20]Available:https://machinelearningmastery.com/bagging-and-random-forest-ensemble-algorithmsfor-machinelearning/