Used care price prediction using machine learning model

CHHEOM SOTHY

Advisor: CHAN Sophal and PHAUK Sokkhey

Institute of technology of Cambodia

Abstract

Used cars are widely sold in the market and online markets. Predicting used car prices is gaining traction in research. Accurately predicting the price of a car is very important as it saves time and boosts sales and purchases. In this article we perform data mining and learn how to use four different models to predict and estimate the value of a used car. Using data of 10 different used cars and a total of more than 100,000 vehicles, the actual forecast is made. Comparing four different linear regression models applied will not yield different results.

Introduction

The pricing of a new car is precise and accurate because the price of a new car is determined by the company itself according to its type, and the use of car parts and accessories does not come with a new car that requires you to purchase. More orders from the company. The price of a new car also varies according to the import tax value of each country, if the car is not manufactured in that country, the sales and some conditions of the government of each country. Therefore, we can clearly determine the price of a new car based on (prices tax and accessories). New cars are usually used a lot in developed countries because developed countries have their own car assembly plants.

After use new car, there are always a lot of used car sales. However, selling used car is very difficult to determine the price because the seller is not experienced in selling used car. On the other hand, in developing countries, imports of used cars are high because developing countries have their own factories, which require buyers to pay import duties, and new car taxes are more expensive than used cars. Buying and selling is valued by the broker. As a result, sellers and buyers of used cars easily fall into the trap of price fraud by brokers.

It is necessary that a system be set up to estimate the value of a used car. To get high-accuracy forecasts, we need to select the best data that comes with the cleanup and select the data that makes the forecast more efficient, using many good features. In this practice we will learn about model machine learning and use the best and most error-free and lowest model to get the best results. We hope that this system will help used car sellers and used car buyers to be more active in selling and buying, which will drive the used car market higher.

II. Literature survey

Used car price forecasts have been extensively studied in various studies. Researchers often predict product prices using some previous data. Puthanut predicts the price of cars in Mauritius, and these are used cars. [1] He used multiple linear regression, k-nearest neighbours, Naïve Bayes and the decision tree algorithm to predict values. A comparison of the predictive results of these techniques shows that the values from these methods are closely comparable. However, it was found that Naïve Bayes decision algorithms and tree methods could not classify and predict numerical values. Puthanut's research also concludes that a limited number of cases do not provide high predictive accuracy [1].

Nitis Monburinon et al. [2] Proposed forecast of prices for used cars using Regression models. In this the paper authors selected data from a German e-commerce site. The main goal of this work is to find appropriate prediction model to predict the price of a used car. They used different machine learning techniques for comparison and used the absolute error mean (Mean absolute error) as a meter. They suggested that their model with the slope promotes regression there is an error below Mean absolute error (0.28) and this gives higher performance where linear regression has Mean absolute error a random forest value of (0.55) with a (Mean absolute error) value of (0.35).

Samerchand Pudaruth [3] proposed to estimate the price of used cars using machine learning techniques. In this paper They collected historical data of used cars in mauritius from the newspaper and applied differently machine learning techniques such as decision tree, K-nearest. Neighbors of multiple linear regression and Naïve Bayes algorithms to predict values. This model makes sense the error is about 27000 for Nissan and about 45000 for toyota uses KNN and about 51000 using linear regression. Accuracy of decision and NaïveBayes algorithm dangled between 60% and 70% with differences the parameters and accuracy of the whole training of the model are 61%.

Zhang et al. [4] use Kaggle data-set to perform price prediction of a used car. The author evaluates the performance of several classification methods (logistic regression, SVM, decision tree, Extra Trees, AdaBoost, random forest) to assess the performance. Among all these models, random forest classifier proves to perform the best for their prediction task. This work uses five features (brand, powerPS, kilometer, sellingTime, VehicleAge) to perform the classification task after removal of irrelevant features and outliers from the dataset which gives an accuracy of 83.08% on the test data. We also use Kaggle data-set to perform prediction of used-car prices. However, the difference lies in the inclusion of few more relevant features in prediction model - the price of the car, and vehicleType. These two features play an important role in predicting the price of a used car which seems to be given less importance in the paper [4]. In addition to this, the range of features year of registration, PowerPS, the price seems to be narrowed down in work [4] due to which test data-set gives less accuracy w.r.t what we evaluate by broadening the range of the above-said features.

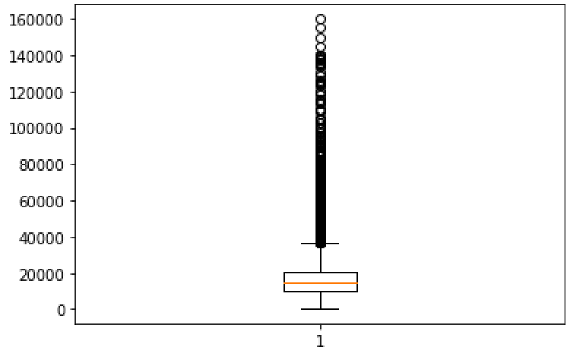
To accurately estimate car prices, different methods are used in the digital world, from machine learning methods such as Multiple linear regression, k-nearest neighbor and naive bayes to random forest and decision tree to SAS Enterprise Explorer. In [5], [6], [7], [8] and [9] all. Of these solutions, different sets of attributes are taken into account when making predictions based on the historical data used to train the model.

In addition, Pudaruth [10] applied various machine learning algorithms, such as: Neighbors nearest, multi-linear regression analysis, decision trees, and naesve bayes for predicting car prices in Mauritius. The data set used to create the forecast model was manually collected from local newspapers in less than a month, as time could have a significant effect on vehicle prices. He studied the following characteristics: brand, model, cubic capacity, distance in kilometers, next year production, color, transmission type and price. However, the authors found that Naive Bayes and Decision Tree could not predict and classify values numerically. In addition, a limited number of data sets cannot provide a high classification, i.e. less than 70% accuracy.

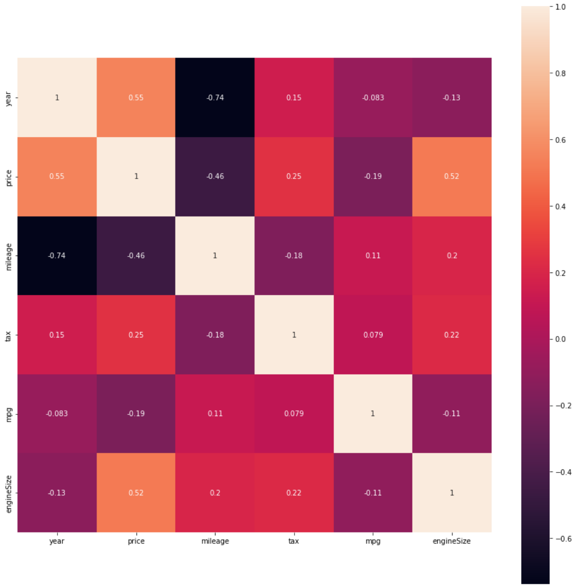
III. Technology

1 Requirements

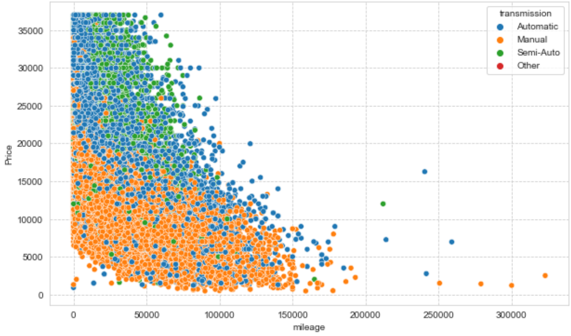
(a) Python as a programming language.
(b) Jupyter as an IDE.
(c) Boxplots.
(d) Correlogram.
(e) scatter plot
(f) Dispersion Plot
(g) Label Encoding



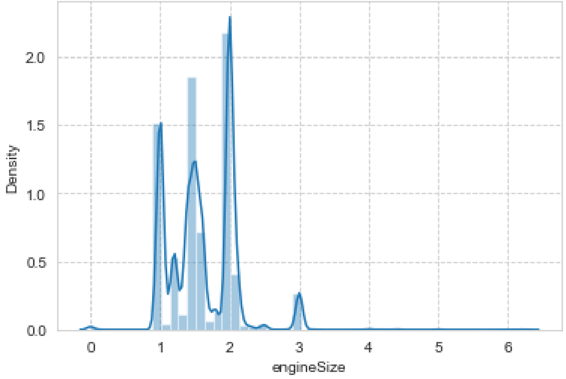
Feature in boxplots of data price



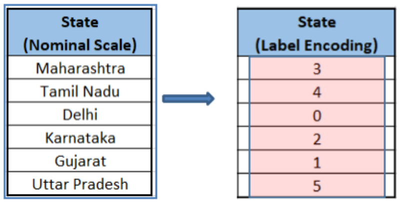
Heat map 6 feature



Scatter Plot by Selecting 3 Features price and mileage to Represent



Dispersion Plot of engine size



State (Nominal Scale)
Maharashtra
Tamil Nadu
Delhi
Karnataka
Gujarat
Uttar Pradesh

State (Label Encoding)
3
4
0
2
1
5

Label Encoding

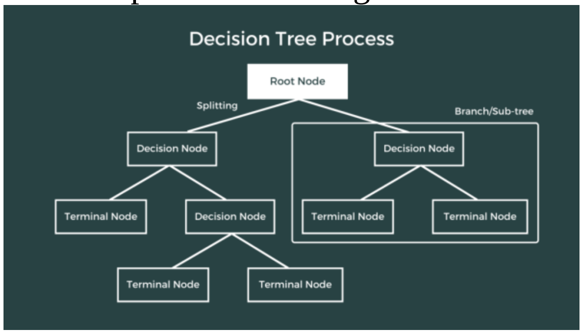
2 Machine learning model

(a) Sklearn for Machine learning.

In machine learning, a kernel refers to a method that allows us to apply linear classifiers to non-linear problems by mapping non-linear data into a higher-dimensional space without the need to visit or understand that higher-dimensional space.

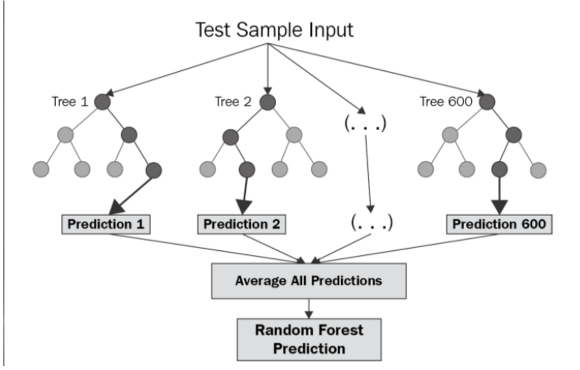
(b) Decision tree regression

Decision tree algorithm creates a tree like conditional control statements to create its model hence it is named as decision tree. Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous.



(c) Random Forest Regressor

Random forest is a meta estimator that fits a number of classifying decision trees on various sub-samples of the dataset and uses averaging to improve the predictive accuracy and control over-fitting.



(d) Bayesian Ridge regressor

Bayesian regression allows a natural mechanism to survive insufficient data or poorly distributed data by formulating linear regression using probability distributors rather than point estimates. The output or response 'y' is assumed to drawn from a probability distribution rather than estimated as a single value.

$$P\left(\mathbf{y}|\theta\right) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma^2}\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)^\top\left(\mathbf{y}-\mathbf{X}\mathbf{w}\right)\right]$$

(e) Linear Regression

Linear Regression is a machine learning algorithm based on supervised learning. It performs a regression task. Regression models a target prediction value based on independent variables.

$$J(\theta) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta(x^{(i)}) - y^{(i)}\right)^2$$

(f) Feature Selection

Backward Elimination s a feature selection technique while building a machine learning model. It is used to remove those features that do not have a significant effect on the dependent variable or prediction of output.

IV. Methodology

This project is divided into 5 subsections as follows:

(a) Data collection

I have used the data of kakle.com in a csv format. Split data into 2 main sets: Training Set for 85% and Testing Set for 15%.

(b) Scaler Data

Using Standard Scaler for scaling the features according to this formula:

$$z = \frac{x - \mu}{\sigma}$$

(c) Feature Selection

Using Backward Eliminations to select only the feature that provide a good model.

(c) Regression Model

Train data with 4 models:

Model 1: Linear Regression

Model 2: Bayesian Ridge regression

Model 3: Decision tree regression

Model 4: A random forest regression

(e) MSE and MAE

Mean squared error (MSE) measures the amount of error in statistical models. It assesses the average squared difference between the observed and predicted values.

$$MSE = \frac{\sum(y_i - \hat{y}_i)^2}{n}$$

The mean absolute error of a model with respect to a test set is the mean of the absolute values of the individual prediction errors on over all instances in the test set.

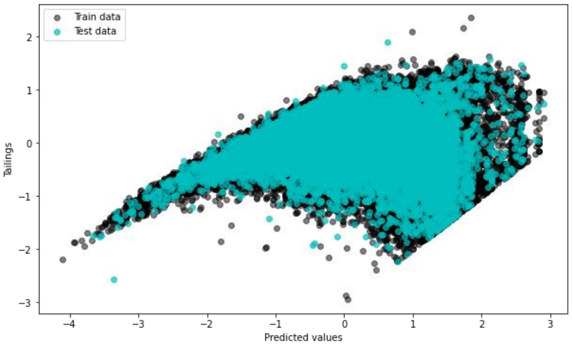
$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|x_i - x|$$

V. Result

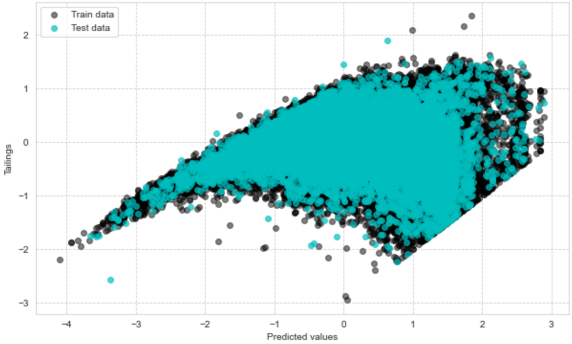
After we train all model and we get one model that has higher score and lowest error is random forest regression model.

Model	Train score	Test score	Mean Squared Error	Mean Absolute Error
01	0.77203	0.77164	0.224830	0.366309
02	0.77203	0.77164	0.224830	0.366309
03	0.96328	0.94310	0.056011	0.166083
04	0.99271	0.95030	0.0489255	0.1504649

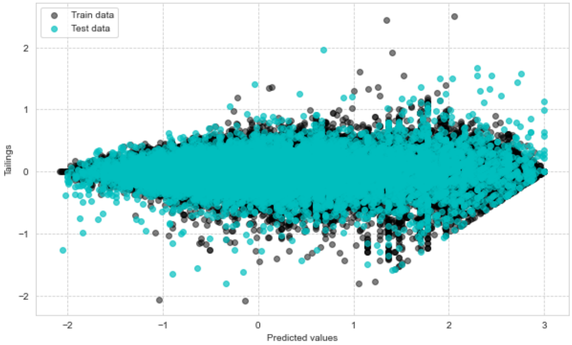
Model 01



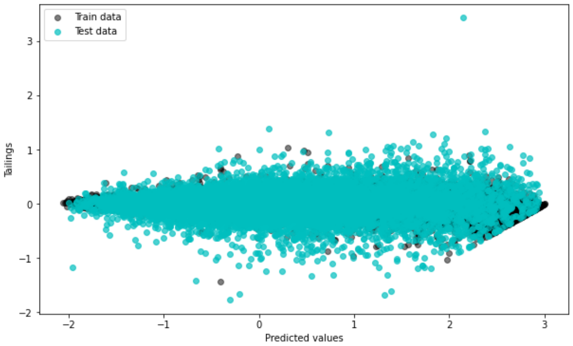
Model 02



Model 03



Model 04



VI. Conclusion and future Work

In the results obtained after the training of the four models, it was observed that the random forest regression worked well. 95% of the results obtained after the test in the training and 15% error on the test on the actual data. We can use the wife to easily estimate the price of a used car.

For the work to be done after the research on this topic, we observe that there are some errors in data preparation that increase MAE up to 15%. For the next study, we will try to reduce MAE.

[1] Pudaruth, S. (2014) ‘Predicting the Price of Used Cars using Machine Learning Techniques’, International Journal of Information & Computation Technology, 4(7), pp. 753–764. Available at: http://www.irphouse.com.

[2] Monburinon, Nitis, Prajak Chertchom, Thongchai Kaewkiriya, Suwat Rungpheung, Sabir Buya, and Pitchayakit Boonpou. "Prediction of prices for used car by using regression models." In 2018 5th International Conference on Business and Industrial Research (ICBIR), pp. 115-119. IEEE, 2018

[3] Pudaruth, Sameerchand. "Predicting the price of used cars using machine learning techniques." Int. J. Inf. Comput. Technol 4, no. 7 (2014): 753-764.

[4] Xinyuan Zhang , Zhiye Zhang and Changtong Qiu, “Model of Predicting the Price Range of Used Car”, 2017

[5] Comparative Analysis of Used Car Price Evaluation Models, Tongji University, Shanghai 200000, China.

[6] Nitis Monburinon, “Prediction of Prices for Used Car by Using Regression Models”, 5th International Conference on Business and Industrial Research, (ICBIR), Bangkok, Thailand, 2018

[7] Jaideep A Muley, “Prediction of Used Cars’ Prices by Using SAS EM”, Oklahoma State University

[8]Nabarun Pal, “A methodology for predicting used cars prices using Random Forest”, Future of Information and Communications Conference, 2018

[9]Kuiper, Shonda, "Introduction to Multiple Regression: How Much Is Your Car Worth?" - Journal Of Statistics Education, 2008.

[10] Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.