Automatic Price Detection of Used Cars Using Machine Learning

Abhishek Raj
Galgotias University
Greater Noida, India
abhishek_raj.scsebtech@galgotiasu niversity.edu.in

Balbindar Kaur
Galgotias University
Greater Noida, India
balbindar.kaur@galgotiasuniversity.edu.in

Abstract

The aim of this study is to develop a model that can predict fair used car prices based on various factors such as vehicle model, year of manufacture, fuel type, price, mileage. In the used car market, this strategy can benefit sellers, buyers and car manufacturers. It can then create a reasonably accurate price estimate based on the data users provide. Machine learning and data science are used in the model building process. The data was taken from used car ads. To achieve maximum accuracy, researchers used various regression approaches, including linear regression, polynomial regression, support vector regression, decision tree regression, and random forest regression. This project visualized the data to better understand the dataset before starting the model building process. To ensure the performance of the regression, the data set was split and transformed to fit the regression. R-squared was used to evaluate the performance of each regression. The final model includes more elements of used cars than earlier research, while also having higher forecast accuracy.

Keywords –

- Analysis, Machine Learning
- Linear Regression
- Ridge Regression ● Lasso Regression
- Random Forest
- XGBoost

Introduction

Determining whether an advertised price is accurate is a difficult task because there are many factors that affect the market price of a used car. The goal of this research is to create a machine learning model that can accurately predict the price of a used car based on its characteristics, allowing buyers to make informed decisions. We build and analyze several learning approaches on top of a dataset containing retail prices for various makes and models. Explore the results of numerous machine learning algorithms. B. Linear Regression, Ridge Regression, Lasso Regression, Elastic Net, Decision Tree Regressor, and Best Fit Selection. Car prices are determined based on several factors. The regression algorithm is used because the output is a sequential number rather than a categorized value. This allows you to predict the actual price of the car rather than its price range. A user interface was also created that takes input from all users and displays the price of the car based on that input. Here are his 3 types of fuel records. Diesel, gasoline and LPG are used here.

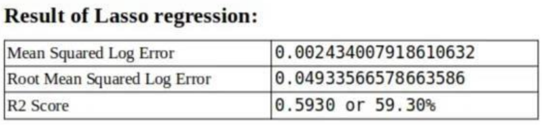
RELATED WORK

Predicting used car prices using machine learning techniques is the first paper. This study investigates how to use supervised machine learning techniques to estimate used car prices in Mauritius. Forecasts are based on historical data from daily newspapers. Various techniques such as multiple regression analysis were used for prediction. According to author Sameerchand, car price estimates are based on historical data from daily newspapers. To estimate the price of the car, they used a supervised machine learning algorithm. Other methods in use include multiple linear regression, k-nearest neighbor algorithms, Nave-based, and various decision tree algorithms. After comparing all four algorithms, the best prediction algorithm was identified. I struggled to compare algorithms, but I succeeded. According to the authors Enis Gegic et al. This white paper focuses on scraping data from online sites using web scraping techniques. We then used various machine learning techniques to compare these and easily predict vehicle prices. They divided the prices into predefined price groups. We developed classifier models using artificial neural networks, support vector machines, and random forest methods on different data sets. In this study, Wu et al. Display Car Price Prediction Using Neural Fuzzy Knowledge-Based SystemThey considered attributes of manufacturer, year of manufacture, and engine type to predict a model that yielded similar results to a simple regression model. Car dealers have a strong desire to sell their leased vehicles at the end of their lease, so they developed an expert system called ODAV (Optimal Distribution of Auction Automobiles). This method will give you information on the best vehicle prices and the best places to get them. We estimated the price of a car using a K-nearest neighbors machine learning approach based on a regression model. With this system, more vehicles are being transferred and therefore managed more effectively. According to author Pattabiraman, the study focuses on the relationship between sellers and buyers. To predict the price of a four-wheeler, we need other characteristics such as: B. Already quoted price, mileage, make, model, equipment, type, cylinders, liters, doors, cruising speed, sound, leather. We used statistical analysis techniques for exploratory data analysis to predict vehicle prices based on these characteristics.

MODULE DESCRIPTION

In this section, we'll go over the many algorithms and datasets that were used to create this module. The model will be trained using a dataset with 92386 records. The value of an automobile is determined by factors such as kilometers travelled, year of registration, fuel type, car model, financial power, car brand, and gear type. We implemented five algorithms because this is a regression problem: Lasso Regression, Ridge Regression, Linear Regression, Random Forest, XGBoost.

A. Lasso Regression The lasso regression allows you to shrink or regularize these coefficients to avoid overfitting and make them work better on different datasets. This type of regression is used when the dataset shows high multi collinearity or when you want to automate variable elimination and feature selection.



Result of Lasso regression:

Mean Squared Log Error	0.002434007918610632
Root Mean Squared Log Error	0.04933566578663586
R2 Score	0.5930 or 59.30%

Fig.1. Final Result of this model

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}|\beta_j|^{\square}$$

Fig.2. Mathematical Formula of Lasso Regression

Where,
Xij = Features of Y or Independent Variable Yi = Dependent Variable

βi = Weights or Magnitude shows importance of a feature λ = minimize the cross-validation prediction error rate.

B. Ridge Regression Ridge regression is a method of estimating the coefficients of multiple-regression models in scenarios where linearly independent variables are highly correlated. It has been used in many fields including econometrics, chemistry, and engineering.Ridge regression is a sort of linear regression that introduces a little degree of bias in order to improve long-term predictions. Ridge regression is a model regularization technique that reduces the model's complexity.L2 regularization is another name for it.The cost function is changed in this method by including a penalty term.

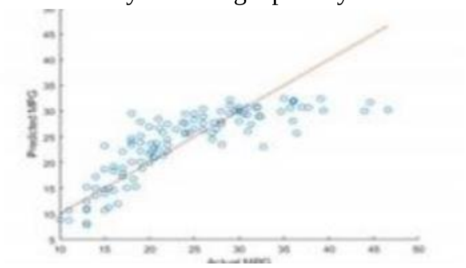


Fig.3.Graphical Representation of Ridge Regression

Ridge Regression penalty is the degree of bias introduced into the model. We may determine it by multiplying the squared weight of each individual label by the lambda.

C. Linear Regression Quick to train and test as a baseline algorithm.

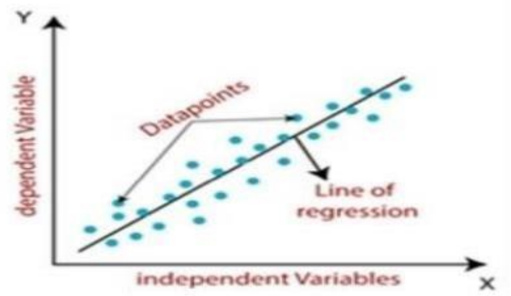


Fig.4.Graphical Representation of Linear Regression

D. The random forest is a classification algorithm consisting of many decision trees. It uses bagging and feature randomness when building each tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

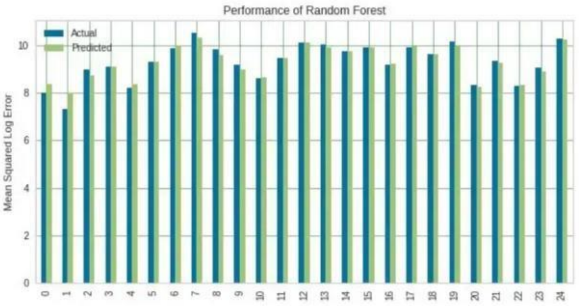
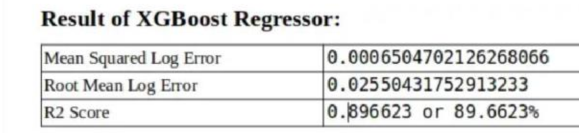


Fig.5. Performance of Random Forest (True value vs predicted value)

E. XGBoost is an ensemble learning method. XGBoost is an implementation of gradient boosted decision trees designed for speed and performance. The beauty of this powerful algorithm lies in its scalability, which drives fast learning through parallel and distributed computing and offers efficient memory usage



Result of XGBoost Regressor:

Mean Squared Log Error	0.0006504702126268066
Root Mean Log Error	0.02550431752913233
R2 Score	0.896623 or 89.6623%

Fig.6. Accuracy Table of XGBoost Regressor

OBJECTIVE

A. Create an efficient and effective model for estimating used car prices based on user input

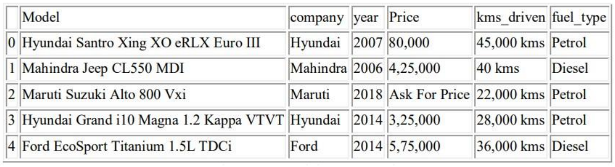
B. Achieve high accuracy**.**

C. To create a user-friendly user interface (UI) that takes input from users and predicts prices**.**

DATA COLLECTION

The data set is specified by column and classified as

1) Company
2) Model
3) Fuel type
4) Kilometers
5) Year of purchase



	Model	company	year	Price	kms_driven	fuel_type
0	Hyundai Santro Xing XO eRLX Euro III	Hyundai	2007	80,000	45,000 kms	Petrol
1	Mahindra Jeep CL550 MDI	Mahindra	2006	4,25,000	40 kms	Diesel
2	Maruti Suzuki Alto 800 Vxi	Maruti	2018	Ask For Price	22,000 kms	Petrol
3	Hyundai Grand i10 Magna 1.2 Kappa VTVT	Hyundai	2014	3,25,000	28,000 kms	Petrol
4	Ford EcoSport Titanium 1.5L TDCi	Ford	2014	5,75,000	36,000 kms	Diesel

Fig6: Processed data set sample

FEASIBILITY

Real-time tracking has already been used in the past, as evidenced by literature reviews (refs 1-5).

- Use a simple answer to the unique problem defined in the problem evolution**.**
- We will employ straightforward things to answer the oneof-a-kind problem defined in the development of the problem.
- Python is used as the programming language. Code is available on GitHub
- Both before and after execution, this project is costeffective. There will be no price for participating in the online surveys. Following the product introduction, it is assessed using survey findings.
- The product management process is completely automated. Our products are created automatically and regularly serviced, and our backend personnel is available 24 hours a day, seven days a week.
- Proper testing and execution take around 2-3 months. The aspects that take the most time are surveys and testing.

CONCLUSION AND FUTURE WORK

Predicting car prices can be a daunting task as there are many features that need to be explored for effective forecasting. Data collection and preparation are the most important steps in the forecasting process. Vehicle data collected from kaggle.com is converted to CSV format and used to create machine learning algorithms during research. Three of his algorithms were used in this study: linear regression, lasso regression and ridge regression. The SVM classifier separated the data into his two parts (support vector machines) for training and testing purposes. That is, 75% of the data was used for machine learning and 25% of the data was used for machine learning. Accuracy of three machine learning models was tested and compared. This is an important comparison between single-group and multi-group machine learning algorithms. As a result, this model helps predict the actual price of the car.

Reference

[1] Enis Gegic, Becir Isakovic, Dino Keco, Zerina Masetic, Jasmin Kevric. “Car Price Prediction Using Machine Learning”;(TEM Journal 2019).

[2] Sameerchand Pudaruth, “Predicting the Price of Used Cars using Machine Learning Techniques”;(IJICT 2014). [3] Richardson, M. S. (2009). Determinants of used vehicle resale value

[4] Wu,etal,(2009). An expert system of price forecasting for used vehicles using adaptive neuro-fuzzy inference.

[5] Doan Van Thai, Luong Ngoc Son, Pham Vu Tien, Nguyen Nhat Anh, Nguyen Thi Ngoc Anh, “Prediction car prices using qualify qualitative data and knowledge-based system” (Hanoi National University)