Machine Learning

Midterm Project Report

Author: Murpys D. Mendez

Project Title

Standardization of Car Prices Using Machine Learning

Introduction

In this project we will use machine learning algorithms to predict car prices. Our objective is to implement a tool to standardize price setting based on physical features and historically established data. Estimating car prices in this way, would set the bases for more objective and fair transactions, and it would provided a unique framework of reference to both buyers and sellers.

The dataset under consideration includes a range of car attributes collected from real-world sources, such as mileage, age, make, model, and engine specifications. Unlike commercial tools, our objective is to establish a reasonable baseline that can be referenced amidst the pull tendency of supply and demand. The results can be cross checked against above mentioned tools (e.g KBB) to have an image of circumstantial price variation vs standard historical figures.

Dataset

Title: Vehicle Sales Data

Source: Kaggle

Description: Vehicle/Car Sales Trends and Pricing Insights

Key Features:

Vehicle Details: Includes specific information about each vehicle, such as its make, model, body style, trim, and manufacturing year.

Transaction Information: Provides insights into the sales transactions, including selling prices and sale dates.

Historical Market Trends: MMR values offer an estimate of the market value of each vehicle, allowing for analysis of market trends and fluctuations.

Condition and Mileage: Contains data on the condition of the vehicles as well as their odometer readings, enabling analysis of how these factors influence selling prices.

Related Work

Kuiper, S. “Introduction to Multiple Regression: How Much Is Your Car Worth?”. *Journal of Statistics Education*, 16(3). 2008 https://doi.org/10.1080/10691898.2008.11889579

Study on several hundreds of GM vehicles from 2005 applying multiple regression.

Mr. Ram Prashath R, et al. “Price Prediction of Used Cars Using Machine Learning”. International Journal for Research in Applied Science & Engineering Technology, Volume 10 Issue V May 2022

Study of different regression approaches including linear , polynomial, support vector machine, decision tree and random forest regression.

Ifthikar, Amjath & Vidanage, Kaneeka. “Valuation of Used Vehicles: A Computational Intelligence Approach”. (2019) 10.1109/ISMS.2018.00011.

GBR, SVM and Naive Bayes used to analyze and predict car data.

Related Work(cont)

Dong Nguyen. “Used car price estimation: 96% accuracy”. 2021 https://www.kaggle.com/code/winternguyen/used-car-price-estimation-96-accuracy?kernelSessionId=63631095

Seeks to filter out unrealistic prices by focusing in essential features.

Mehar Vijh, et al. "Stock Closing Price Prediction using Machine Learning Techniques". Procedia Computer Science, Volume 167, 2020, Pages 599-606, ISSN 1877-0509,https://doi.org/10.1016/j.procs.2020.03.326.

Uses Artificial NN and random Forest to predict prices.

C. Jin, "Price Prediction of Used Cars Using Machine Learning," IEEE International Conference on Emergency Science and Information Technology (ICESIT), Chongqing, China, 2021, pp. 223-230, doi: 10.1109/ICESIT53460.2021.9696839.

Application of Various regression algorithms *and the use of R square to evaluate performance.*

Methodology

Our study will comprise of four basic stages that will be implemented in R Studio.

1-Feature selection. As the problem is defined as “prediction of car prices based on physical attributes”, we will subject the features to a screening in order to discard those that have no, or negligible weight on the outcome variable.

2-Data cleaning and normalization . We will remove missing values and make sure the encoding is consistent throughout the data and it has a suitable format for processing by our ML algorithms.

3-Model selection. In an initial approach, we intend to explore a few machine learning algorithms for regression, including linear regression, decision trees and random forest, applied to our training data. The data will be split into training and test sets with a ratio of 80-20%

4-Performance evaluation. Finally we will conclude on which model performed best based on standard statistical metrics.

Experiments

- We will start by cleaning the data and selecting features that have greater impact on the target variable.
- We will then study the relationship between features and the outcome variable, if the relationship is linear we will train and evaluate a linear regression model.
- Depending on the results, we might train other supervised methods, like decision trees, random forest or SVM.
- We will verify how feature selection impact each model.
- Comparison of the results using R square and RMSE.

Timeline

Week 10: Perform EDA. Determine relationship between features and target variable.

Week 11: Begin R code with linear regression if applicable.

Week 12: Explore decision tree/random forest.

Week 13: Consider SVM. Choose best algorithm.

Week 14: Organize code into a report format.

Week 15: Finish and deliver project.

References

Maulud, Dastan, and Mohsin Abdulazeez, Adnan. “ Review on Linear Regression Comprehensive in Machine Learning”. Journal of Applied Science and Technology Trends, vol. 1, no. 2, Dec. 2020, pp. 140-7

S. Pathak, I. Mishra and A. Swetapadma. "An Assessment of Decision Tree based Classification and Regression Algorithms". *2018 3rd International Conference on Inventive Computation Technologies (ICICT)*, Coimbatore, India, 2018, pp. 92-95, doi: 10.1109/ICICT43934.2018.9034296.

Hankar, M., Birjali, M., Beni-Hssane, A. “Machine Learning Modeling to Estimate Used Car Prices.” Innovations in Smart Cities Applications Volume 6. SCA 2022. Lecture Notes in Networks and Systems, vol 629. Springer, Cham. (2023).https://doi.org/10.1007/978-3-031-26852-6_49

S. Shaprapawad, P. Borugadda and N. Koshika, "Car Price Prediction:An Application of Machine Learning," 2023 *International Conference on Inventive Computation Technologies (ICICT)*, Lalitpur, Nepal, 2023, pp. 242-248, doi: 10.1109/ICICT57646.2023.10134142.

Claudio Gambella, et al. "Optimization problems for machine learning: A survey". European Journal of Operational Research. Volume 290, Issue 3. 2021. Pages 807-828, ISSN 0377-2217. https://doi.org/10.1016/j.ejor.2020.08.045.