Milestone I Report - Problem Identification and Project Specification
Predicting Used Car Prices using Machine Learning

CSCE 3602 - Fundamentals of Machine Learning

Omar Bahgat
Computer Engineering Department
The American University in Cairo
Cairo, Egypt
omar_bahgat@aucegypt.edu

Omar Saleh
Computer Engineering Department
The American University in Cairo
Cairo, Egypt
Omar_Anwar@aucegypt.edu

Introduction

The automotive industry has witnessed a significant increase in the availability of data regarding various features of cars such as brand, model, and production year. This huge amount of data presents an opportunity to leverage machine learning techniques to be able to predict car prices accurately.

The ability to predict car prices is of utmost importance for various stakeholders, including buyers, sellers, manufacturers, and dealerships. Buyers and sellers seek fair pricing when buying/selling a car, while manufacturers and dealerships aim to optimize their pricing strategies to remain competitive in the market.

The problem at hand involves developing a machine-learning model capable of accurately predicting car prices based on a diverse set of features. This project aims to address this challenge by employing supervised machine learning algorithms we will be covering in class on a dataset consisting of cars and their features and choosing the most optimal one.

Literature Review

There have been several attempts to solve the problem of predicting used car prices using machine learning techniques. We will be going briefly over three different papers analyzing their attempts to solve the problem, the machine learning approach they used, and the performance of their model.

In the first paper, “Prediction of Used Car Prices Using Machine Learning” by Dibya Ranjan et.al.[1], the authors used different machine learning algorithms such as k-nearest neighbor (KNN), random forest regression, decision tree, and light gradient boosting machine (LightGBM) to predict the price of used cars based on different features specific to Indian buyers. They implemented the best model by comparing it with other models. Implementing LightGBM regression, which had the least RMSE value along with the least time taken to train and test the model.

In a second paper “Car Price Prediction using Machine Learning Techniques” by Enis Gegic et al., published in the TEM Journal in 2019 [2], the authors conducted an in-depth study on predicting the price of used cars in Bosnia and Herzegovina. They applied three different machine learning algorithms: Artificial Neural Network, Support Vector Machine, and Random Forest. These algorithms were used in combination to build a robust model for predicting the price of used cars.The data used for the prediction was collected from the web portal autopijaca.ba.This data included various features of the cars, which were then used as input for the machine learning algorithms. The final prediction model was integrated into a Java application. Furthermore, the model was evaluated with an accuracy of 87.38% The authors highlighted the potential of machine learning in predicting used car prices. However, they indicated that each approach has its strengths and weaknesses, and the choice of model can depend on

various factors, such as the characteristics of the dataset and the computational resources available.

In a third paper “Forecasting resale value of the car: Evaluating the proficiency under the impact of machine learning model” by Harshit Gupta et al. [3], the authors present a new approach to forecasting car prices using multiple gradient boosted decision trees to minimize overfitting and eliminate outliers and irregularities from the predictions. The machine learning model was trained using a Ten-fold cross-validation approach where the data points were randomly assigned to ten partitions and nine of those partitions were used for training while the last partition was used for validation. This process was repeated ten times where each of the partitions was used once for validation. From the model results, Gupta concluded that the major features affecting car prices were the car manufacturer, the distance driven, the fuel type, and the transmission configuration. The machine learning model was evaluated and it achieved an accuracy of 93.73% on a test set.

Project Overview and Applications

In this project, our goal is to build a machine learning model for predicting used car prices accurately. We aim to use a dataset comprising various features of cars, including make, model, year, mileage, condition, and several others. We will start by preparing our dataset to make it ready for usage by different machine learning models. Our approach will involve cleaning the data to handle missing values and filtering irrelevant features and pre-processing the data to encode categorical variables and scale numerical features. Afterwards, we plan to test with all possible supervised machine learning algorithms discussed in class to determine the most optimal one for our dataset. Finally, we will design our final model and pick an algorithm such that it best fits our data and objectives and yields the best performance.

Overall, our proposed solution aims to offer valuable insights into the dynamics of used car pricing and can significantly benefit various stakeholders in the automotive industry. For example, buyers can make informed decisions and purchase vehicles based on fair market values. On the other hand, sellers can utilize accurate price predictions to set competitive yet profitable prices for their vehicle. Car dealerships can optimize their inventory management through ensuring they stock the right vehicles at the right prices to meet demand effectively. Additionally, car manufacturers can leverage price prediction models to gain insights into consumer preferences and market trends.

Dataset I (Main)

The dataset chosen for our project, authored by Andrei Novikov, contains data on around 760,000 used cars which were scrapped from cars.com during April 2023.

With 20 distinct features, the dataset covers everything from basic car information like car manufacturer, model, and production year, to specifics like mileage, engine type, and transmission configuration. It also includes details like fuel type, MPG (miles per gallon), exterior and interior colors, accident history, and ownership status.

The main point of our analysis lies in the "Price" feature, serving as the label for our model. This feature signifies the listed price of each used car on cars.com which is the output our machine learning model will try to predict accurately.

While the dataset has a huge range of used cars and a diverse set of features, some of the features contain missing values. For example, MPG has a 19% missing value rate and interior color has a 7% missing value rate. However, those missing value rates are considerably good in comparison with other datasets, which influenced our decision to use this particular dataset.

This dataset can be found on Kaggle here [4].

Dataset II

A second potential dataset for our project, compiled by Mikhail Pustovalov, consists of data on

approximately 330,000 used cars collected from auto.ru. This data spans all sales offers available in Russia as of March 15, 2021.

With a total of 32 features, it offers a comprehensive overview of each vehicle, ranging from fundamental details like brand, model, and year to more specific factors such as mileage, engine specifications, and transmission type. Additionally, the dataset includes information on the car's body type, color, fuel type, ownership history, technical inspection status, and more.

Of particular interest to our project is the prediction of the "Price" feature, which serves as the label for our machine learning model.

Although this dataset has plenty of features and a substantial amount of used cars, there are three features with very high missing value rates (77%, 45%, and 21% respectively) which would cause trouble during pre-processing and might affect data distributions negatively if the missing values are not replaced correctly.

This dataset can be found on Kaggle here [5].

Dataset III

A third potential dataset comprises vehicle listings from Craigslist.org, offering a comprehensive collection of used vehicles for sale across the United States. The dataset encompasses approximately 430,000 samples of used car listings, collected over several years up to 2021. Authored by Austin Reese, it includes various features such as price, year, make, model, condition, odometer reading, transmission type, and more with a total of 26 features providing extensive information on each vehicle listing.

For our project, the “Price” feature as previously highlighted could serve as the label for our machine learning model which aims to predict the price of used cars accurately.

While the dataset provides a comprehensive collection of used vehicle listings from Craigslist, it may have limitations in terms of data quality, completeness, and accuracy. Some listings may contain missing or inconsistent information, and there could be biases inherent in the data collection process. Additionally, since the dataset is specific to Craigslist listings, it may not capture the entire used vehicle market comprehensively.

This dataset can be found on Kaggle here [6].

References

[1] Adhikary, D.R., Sahu, R. & Panda, S.P. (2022). Prediction of Used Car Prices Using Machine Learning. In: Dash, S., Das, S., Panigrahi, B.K., Das, S. (eds) Biologically Inspired Techniques in Many Criteria Decision Making. SIST, vol 271. Springer, Singapore https://link.springer.com/chapter/10.1007/978-981-16-8739-6_11

[2] Gegic, E., Isakovic, B., Keco, D., Masetic, Z., & Kevric, J. (2019). Car Price Prediction using Machine Learning Techniques. TEM Journal, 8(1), 113-1181. TEMJournalFebruary2019_113_118.pdf

[3] Arora, P., Gupta, H., & Singh, A. (2022). Forecasting resale value of the car: Evaluating the proficiency under the impact of machine learning model. In: International conference on design and applications of multifunctional materials, interfaces and composites (DAM2IC 2022). Materials Today: Proceedings, vol 69, pp 441–445 https://doi.org/10.1016/j.matpr.2022.09.074

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

[5] Pustovalov, Mikhail. 2021. Auto.ru: All Used Car Offers as of 17.03.2021. Kaggle Available at: https://www.kaggle.com/datasets/mikhailpustovalov/autoru-all-used-car-offers-as-of-17032021.

[6] Reese, Austin. Craigslist Used Cars Dataset. Kaggle. Available at: https://www.kaggle.com/datasets/austinreese/craigslist-carstrucks-data