Used Car Price Prediction Analysis

November 2023

Muma College of Business
University of South Florida

2023

Submitted by :
Mr. Faizan Mian
Mr. Sankalp Thota
Mr. Sonu Kanna
Mr. Yashwanth Danda

Executive Summary

The global health crisis ushered in a new era for motorists as they shifted towards less pricey used cars. As a response, we created a prediction model of used-car prices in the United States. The accuracy of this tool is based on factors such as make, model, mileage, and year. It allows sellers to give competitive prices and enables the buyers to have informed choices.

The report goes deep into the data describing attributes' importance and how it relates to our pricing questions. As a transparent and truthful method for predicting vehicle prices, we have opted for two models, namely, Decision Tree Regressor and Random Forest Regressor.

Our models perform decently in predicting car prices. We've confirmed their accuracy using factors such as MSE, RMSE, MAE, and R2-scores. Furthermore, visualization shows efficacy of the models, depicting actual against forecasted prices.

We do not only analyse the data, but make informed decisions amidst a constantly changing used car industry. This makes it easy for industry players to know all issues that lead to pricing that helps them to form better pricing strategies and competitive advantage with other players in the marketplace.

Introduction

In 2022, the automotive manufacturing industry ranked 9th among the world's top 10 largest industries. Over the previous decade, from 2010 to 2019, the U.S. automotive sector enjoyed steady growth with an impressive annual sales increase of nearly 5%.

However, the outbreak of the COVID-19 pandemic brought about substantial disruptions to this industry. As a result of the pandemic, consumer preferences have shifted towards private transportation, with people seeking safer and more isolated travel options. Yet, the financial challenges posed by the pandemic have created obstacles for new vehicle purchases. Many commuters, facing economic concerns, are turning to the second hand car market as a more budget-friendly alternative.

The model under development is designed to predict the price of used cars, leveraging a comprehensive set of features that play a crucial role in determining the value of these vehicles.

The question we try to resolve with this project is whether we can accurately predict the selling price of used cars in the United States based on various features like manufacturer, model, mileage, year? This is crucial because it can assist both sellers and buyers in making informed decisions. Sellers can price their used cars competitively, and buyers can assess whether the asking price is reasonable. Accurate pricing predictions contribute to a more transparent and efficient used car market, reducing information asymmetry.

Data Characteristics

The following are the independent variables that are used to predict the price of a vehicle:

Manufacturer: This attribute represents the company that produced the vehicle. It can be important for understanding brand reputation and its impact on used car prices.

Model: The specific model of the vehicle, which can affect pricing, as different models may have different market values and demand.

Category: Describes the vehicle's category, such as sedan, SUV, or sports car. The category can influence pricing and depreciation rates.

Fuel type: Specifies the type of fuel the vehicle uses, which can be relevant for pricing, as fuel efficiency impacts operating costs.

Gear box type: Indicates the type of transmission, such as automatic or manual. This can impact pricing and appeal to buyers.

Drive wheels: Describes the drive configuration, which can influence vehicle performance and appeal to certain buyers.

Production year: Represents the production year of the vehicle. This is crucial for assessing depreciation rates and understanding how vehicle age impacts prices.

Mileage: The distance the vehicle has travelled in kilometres, a key factor in determining depreciation rates and used car prices.

Cylinders: Indicates the number of cylinders in the engine, which can impact performance and fuel efficiency, thereby influencing pricing.

Airbags: Describes the number of airbags in the vehicle, which is essential for safety and can impact buyer decisions.

The data attributes, such as Manufacturer, Model, Category, Fuel type, and others, are crucial for understanding the factors that influence used car prices, depreciation rates, and market trends. These attributes provide the information needed to address the questions related to predicting used car prices, identifying factors affecting depreciation rates, and finding models that retain their value.

Train and Test Split:

To build and evaluate a predictive model, we split the dataset into a training set and a testing set. The training set is used to train the model, while the testing set is used to evaluate its performance. We randomly split the data into these sets, with a common split ratio of 70% for training and 30% for testing. This split allows for the model to be trained on a portion of the data and tested on unseen data to assess its predictive accuracy.

Model Construction

The ML models which we have chosen are Decision Tree Regression and Random Forest Regression. The choice of using Decision Tree Regressor and Random Forest Regressor aligns with the data characteristics and problem domain of car price prediction. Given that the target variable is a continuous numerical value (car price), regression models are appropriate for this problem. The need for interpretability is crucial in the context of car price prediction because stakeholders often seek an understanding of the factors influencing prices. A Decision Tree Regressor offers transparency in decision-making, allowing for the inspection of decision paths and feature importance. Furthermore, the subsequent use of Random Forest Regressor is advantageous due to its ability to improve predictive accuracy while retaining a degree of interpretability. This ensemble model overcomes potential overfitting and enhances model robustness, making it well-suited for real-world car price prediction applications.

Evaluation Metrics and Model Performance:

Evaluation metrics play a vital role in assessing model performance. The choice of mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), and R2-score is appropriate for regression tasks like car price prediction. These metrics collectively provide a comprehensive evaluation of your models. MSE and RMSE measure prediction accuracy by assessing the squared differences between actual and predicted values. The issue however with using MSE and RMSE, in the context of the used dataset, is that the dataset contains many outliers. MSE and RMSE are sensitive to datasets with many outliers. Therefore, MAE is a better evaluation metric, since it quantifies how close predictions are to actual values. R2-score was chosen because it indicates the model's ability to explain the variance in car prices. Additionally, we decided to use scatter plots to visualize actual vs. predicted prices, since it aids in understanding the alignment of model predictions with the ideal y=x line. This comparison with the baseline model further assesses the improvement achieved by our regression models, enhancing the overall evaluation of model performance in the context of car price prediction.

Results

Using the Decision Tree Regressor, the model got an MAE value of 5239, meaning that for each car price, the error is either +5239 or -5239 of the actual price. This is not a particularly good score, since an error of $5000+ does not give a good estimate of a car price. Furthermore, the R2 score for the Decision Tree Regressor was 0.59, showing that the data does not fit the model particularly well. To visualize the predicted values vs. the actual values, the regression plot shown in Figure 1 was generated. As can be seen in Figure 1, many predicted points, lie close to the ideal y=x line. However, there are a significant number of "failed prediction" that can be seen, presumably due to the many outliers in the data set.

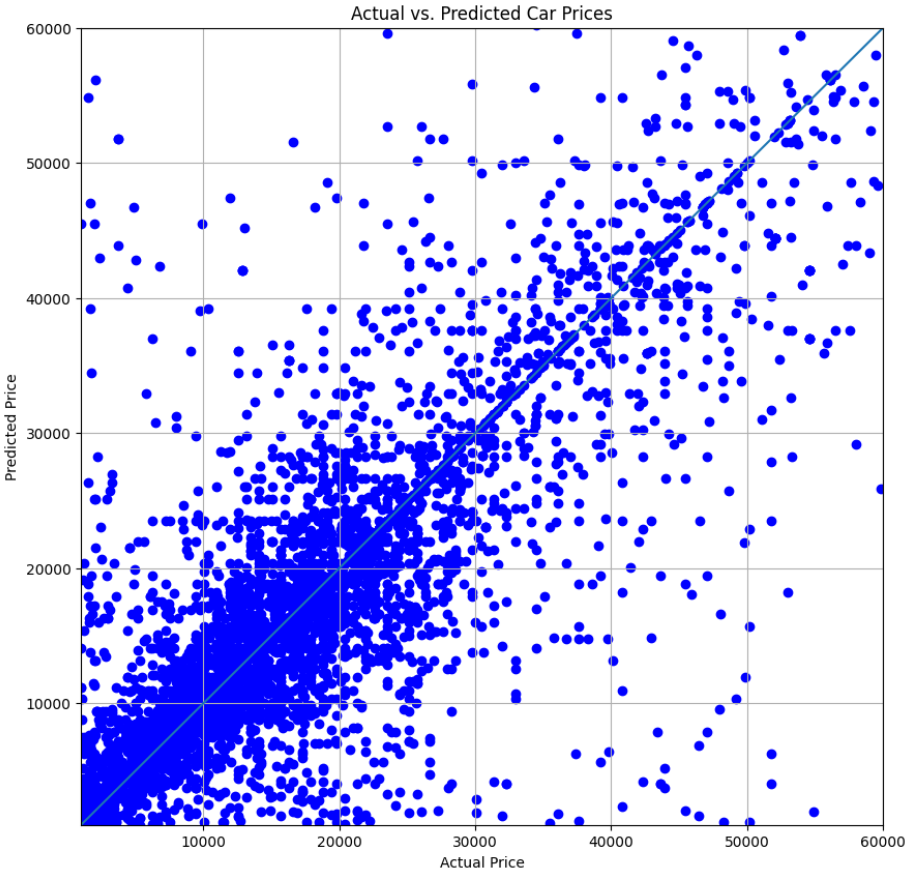

Figure 1: Regression plot for the Decision Tree Regressor in the interval $10000 - $60000

To further advance and improve the Decision Tree Regressor, the ensemble model called Random Forest Regressor was used. The goal with the implementation of the Random Forest Regressor was not to give a near perfect prediction, but rather to improve the predictions that the Decision Tree Regressor made. The Random Forest Regressor got an MAE value of 4484. Such an error is still not to be considered a "good" prediction since an error of $4000+

for a car price is not a good estimate. Although the MAE remains fairly large, it is still an improvement from the previous model. Furthermore, the R2 score for the Random Forest Regressor was 0.70, showing an improvement in how well the data fits the model. To further visualize the improvement when switching from Decision Tree Regressor to Random Forest Regressor, the regression plot shown in Figure 2 was generated. As can be seen from the regression plot below, the model managed to decrease the number of "failed predictions", i.e. the values are not "scattered" around as much as it was in Figure 1.

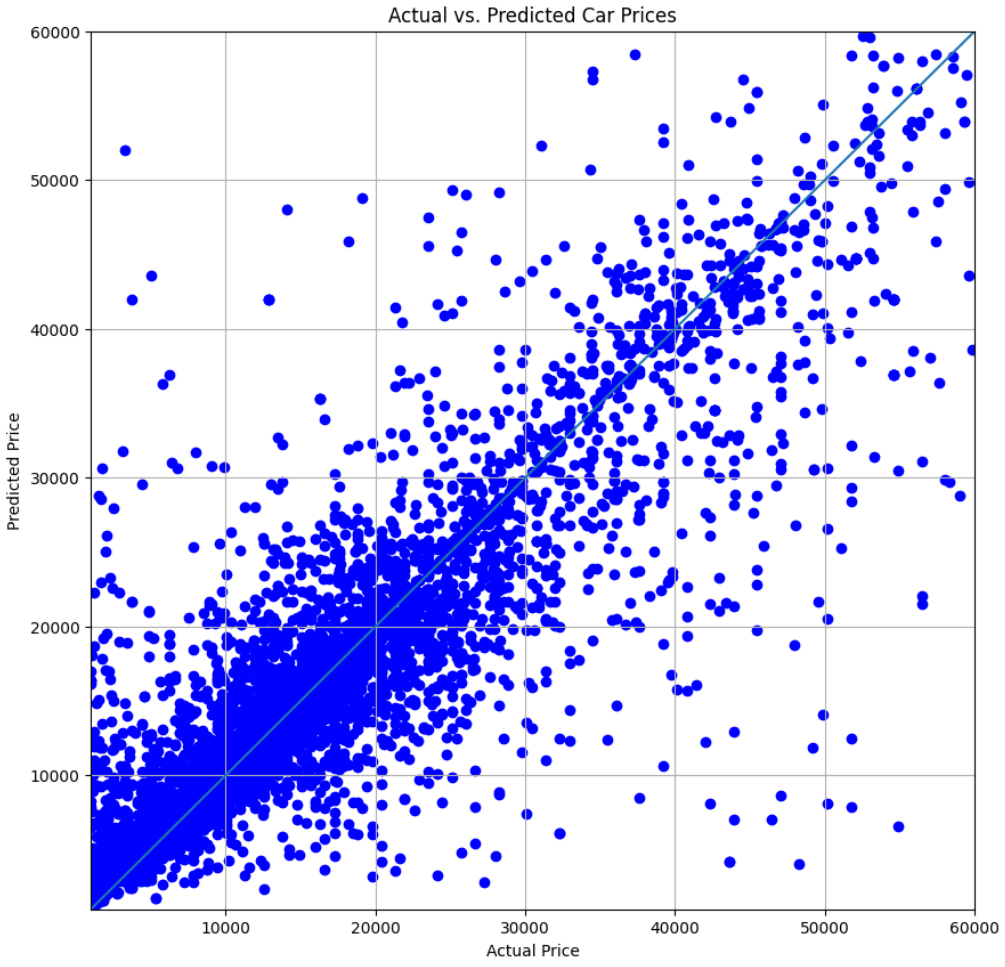

Figure 2: Regression plot for the Random Forest Regressor in the interval $10000 - $60000

Although we managed to improve the performance of the model by implementing Random Forest Regressor, there is still a lot of room for improvement. For instance, in order to make a somewhat good price prediction, the MAE should stay below $500. The main reason for the high MAE scores is the quality of the dataset that was used to train the model. The dataset that was used contained multiple outliers, not only in terms of price, but also in terms of mileage. For instance, some cars had a price of over 20 million dollars, even though they were not supercars and some cars had a mileage well over the average mileage of a car before it is scrapped. Such outliers contribute to a significant deviation from the true line of best fit, leading to inaccurate predictions.