A Statistical Approach to Used Car Price Prediction

Steven Qie(qie2), William Yeh(wy16), Brian Gong(brianhg2)

2024-12-09

Introduction

With the used car market being significantly larger than the new car market, many consumers are realizing that used cars provide a more affordable option. It plays a significant role in the growth and stability of the U.S. economy, driven by changing consumer preferences, economic factors, and the availability of certain cars. Accurately predicting the price of a used car is a challenging but essential task for buyers, sellers, and market analysts/economists alike.

This report aims to develop various predictive models for used car prices using the Used Car Price Prediction Dataset from Kaggle. This dataset comprises of 4,009 data points, representing unique vehicle listings, as well as nine distinct features that serve as key indicators influencing the value of a used car. We follow a very structured and standard approach, including data exploration, preprocessing, model training, and evaluation using relevant performance metrics. By leveraging these methods, we aim to uncover valuable insights into the world of automobiles and the various factors that are driving used car prices.

Need a section on key findings....

We utilized AI tools in this report to enhance and assist in our writing. These tools helped play a big role in ensuring clarity, conciseness, and professionalism. We also utilized AI tools to help us with syntax help when writing code in R, as well as discovering potential bugs in our code.

white space

white space

white space

white space

white space

white space

white space

white space

white space

white space

white space

white space

white space

Literature Review

This literature review aims to summarize the main findings and approaches from a few noteworthy research papers focused on used car price prediction.

The first paper of interest is "Price Prediction of Used Cars Using Machine Learning", written by Chuyang Jin of the University of Sydney. In his paper, he presented a model that can predict a used vehicle's price given certain conditions, as well as abilities to depict used vehicles' depreciation over the years. He hopes that his model can benefit and save time for both sellers and buyers selling or searching for second-hand vehicles in the market. By being able to depict the depreciation of used vehicles over the years, he hopes to help consumers decide which model to purchase if they plan to sell it in the future, as well as assist car manufacturers in determining which models should be produced more. Jin used a CSV dataset containing 100,000 records of used cars in the UK, focusing specifically on the Mercedes brand. The nine factors that he considered were the following: model, year, selling price, transmission, mileage, fuel type, tax, miles per gallon (mpg), and engine size. Notably, several problems started appearing immediately. The distribution of the various indicators was skewed. For example, the overwhelming majority of price falls in the 0-75,000 range, so there were not enough samples of used cars priced over 75,000 to be able to build an effective model. Jin tried various forms of regression, namely linear, polynomial, SVR, Decision Trees, and Random Forests, and discoevered that Random Forest Regression had the best R squared value.

Data Processing and Summary Statistics

Unsupervised Learning

Prediction Models

Open-Ended Question/Conclusion