A Statistical Approach to Used Car Price Prediction

12/09/2024

Steven Qie
Statistics and Computer Science
University of Illinois Urbana-Champaign
qie2@illinois.edu

Brian Gong
Statistics and Computer Science
University of Illinois Urbana-Champaign
brianhg2@illinois.edu

William Yeh
Statistics and Computer Science
University of Illinois Urbana-Champaign
wy16@illinois.edu

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....

Abstract—

white space

white space

white space

white space

white space

white space

white space

white space

white space

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.

Literature Review

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

"Price Prediction of Used Cars Using Machine Learning", written by Chuyang Jin of the University of Sydney, presents a model that can predict a used vehicle's price given their year of production, mileage, tax, miles per gallon, He hopes that his model can benefit and save time for both sellers and buyers who are looking to sell or serach for second-hand vehicles. 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. While doing exploratory data analysis and preprocessing, Jin noted that many many predictors had skewed distributions. For example, the overwhelming majority of prices fell in the 0-75,000 range, limiting the model's potential effectiveness for higher price ranges. Jin deemed these data points as outliers and excluded them to ensure that the model would be more accurate and usable. After testing various forms of regression, namely linear, polynomial, SVR, Decision Trees, and Random Forests, Jin found Random Forest Regression yielded the best R squared value of 0.90416.

"Used Car Price Prediction using Machine Learning: A Case Study", written by Mustapha Hankar, Marouane Birjali, and Abderrahim Beni-Hssane, applies several supervised machine learning algorithms to predict used car price prices based on features from a dataset collected from an online eCommerce website called Avito. During preprocessing, the authors of this paper performed recursive feature elimination to maintain only the most relevant features to car prices: year of manufacture, mileage, mark, fuel type, fiscal power, and model. Along with a baseline multiple linear regression model, the study also looked at K-nearest neighbors, Random Forest, Gradient Boosting, and Artificial Neural Networks. The study utilized 2 different performance metrics, R^2 and RMSE, and concluded that the Gradient Boosting Regression Model achieved the best results, with a R^2 of 0.8 and RMSE of 44516.20.

"Car Price Prediction using Supervised and Unsupervised Learning Models and Deep Learning" by Thomas Nsiah approached the problem of car price prediction from a supervised and unsupervised lenses. While supervised models allow a consumer to understand the key factors and predictors that influence pricing of used cars, unsupervised learning oftentimes uncovers hidden connections and patterns within the data. In his paper, Nsiah used a mock dataset of 50,000 UK second hand car sales with features similar to the previous 2 studies, such as model, engine size, fuel type, year, and mileage. Supervised learning models that Nsiah tried included simple linear regression, polynomial regression, and random forest, evaluated using mean absolute error (MAE) and R-squared metrics. He concluded that out of the supervised models, random forest performed best with an R-squared of 0.99849 and a MAE of 289.0691. For unsupervised learning techniques, Nsiah applied K-Means and DBSCAN clustering to identify price patterns, evaluated using the Davis Boudlin Index and the Silhouette Coefficient. He concluded that K-Means clustering for the year of manufacture vs price produced the best clustering results.

Overall, these three studies demonstrate the effectiveness that machinie learning can have on accurately predicting used car prices. The next section will outline our own approach and findings.

Citations:

- C. Jin, "Price Prediction of Used Cars Using Machine Learning," in 2021 IEEE International Conference on Emergency Science and Information Technology (ICESIT), Chongqing, China, 2021, pp. 223-230, doi: 10.1109/ICESIT53460.2021.9696839.
- M. Hankar, M. Birjali, and A. Beni-Hssane, "Used Car Price Prediction using Machine Learning: A Case Study," in 2022 11th International Symposium on Signal, Image, Video and Communications (ISIVC), El Jadida, Morocco, 2022, pp. 1-4, doi: 10.1109/ISIVC54825.2022.9800719.
- T. Nsiah, "Car Price Prediction using Supervised and Unsupervised Learning Models and Deep Learning," unpublished, 2024.

Data Processing and Summary Statistics

Import data

```
data <- read.csv("used_cars.csv")
```

quick overview of data

```
head(data)
```

```
##      brand                           model model_year      milage     fuel_type
## 1     Ford Utility Police Interceptor Base       2013  51,000 mi. E85 Flex Fuel
## 2  Hyundai                     Palisade SEL       2021  34,742 mi.      Gasoline
## 3    Lexus                   RX 350 RX 350       2022  22,372 mi.      Gasoline
## 4 INFINITI                Q50 Hybrid Sport       2015  88,900 mi.        Hybrid
## 5     Audi       Q3 45 S line Premium Plus       2021   9,835 mi.      Gasoline
## 6    Acura                         ILX 2.4L       2016 136,397 mi.      Gasoline
##                                                 engine      transmission
## 1 300.0HP 3.7L V6 Cylinder Engine Flex Fuel Capability       6-Speed A/T
## 2                                 3.8L V6 24V GDI DOHC 8-Speed Automatic
## 3                                      3.5 Liter DOHC         Automatic
## 4  354.0HP 3.5L V6 Cylinder Engine Gas/Electric Hybrid       7-Speed A/T
## 5                           2.0L I4 16V GDI DOHC Turbo 8-Speed Automatic
## 6                                          2.4 Liter                 F
##                  ext_col int_col                              accident
## 1                  Black   Black At least 1 accident or damage reported
## 2        Moonlight Cloud    Gray At least 1 accident or damage reported
## 3                   Blue   Black                             None reported
## 4                  Black   Black                             None reported
## 5 Glacier White Metallic   Black                             None reported
## 6                 Silver  Ebony.                             None reported
##   clean_title   price
## 1         Yes $10,300
## 2         Yes $38,005
## 3             $54,598
## 4         Yes $15,500
## 5             $34,999
## 6             $14,798
```

```
na_values <- data[!complete.cases(data), ]
cat("NA Values:", nrow(na_values), "\n")
```

```
## NA Values: 0
```

```
empty_space_rows <-  data[rowSums(data == "", na.rm = TRUE) > 0, ]
cat("Empty Values:", nrow(empty_space_rows), "\n")
```

```
## Empty Values: 740
```

Just looking at the data, some notable columns to do some preprocessing is accident,engine. Price and mileage can use some preprocessing to make into a number. Also, milage is spelled wrong. Although there seems to be no NA values, a lot of rows of clean_title are notably empty. Other cols like fuel_type and accident also contain some empty values

```r
cat(unique(data$accident), "\n")
```

```
## At least 1 accident or damage reported None reported
```

```r
data$accident <- ifelse(data$accident == "At least 1 accident or damage reported", 1, 0)
```

Because accident only has 2 unique values, no accidents and 1 or more accidents, we changed it to 1,0 to be useful for models

```r
head(data$engine, 5)
```

```
## [1] "300.0HP 3.7L V6 Cylinder Engine Flex Fuel Capability"
## [2] "3.8L V6 24V GDI DOHC"
## [3] "3.5 Liter DOHC"
## [4] "354.0HP 3.5L V6 Cylinder Engine Gas/Electric Hybrid"
## [5] "2.0L I4 16V GDI DOHC Turbo"
```

```r
library(stringr)
```

```
## Warning: package 'stringr' was built under R version 4.2.3
```

```r
# Extract Horsepower (HP)
data$horsepower <- as.numeric(str_extract(data$engine, "\\d+\\.\\d+(?=HP)"))

# Extract Displacement
data$displacement <- as.numeric(str_extract(data$engine, "\\d+\\.\\d+(?=L)"))

# Extract Cylinders
data$cylinders <- str_extract(data$engine, "\\d+ Cylinder")

# Extract Engine Type
data$engine_type <- str_extract(data$engine, "DOHC|SOHC|Turbo|Twin Turbo|Electric Motor")

# Extract Fuel Type
data$fuel_type <- str_extract(data$engine, "Gasoline|Diesel|Electric|Hybrid|Flex Fuel|Plug-In Electric/
```

```r
cat(head(data$price, 5), "\n")
```

```
## $10,300 $38,005 $54,598 $15,500 $34,999
```

```r
data$price <- as.numeric(gsub("[$,]", "", data$price))
```

Removed the dollar sign and comma in price to enable numeric operations

```r
colnames(data)[colnames(data) == "milage"] <- "mileage"
data$mileage <- as.numeric(gsub("[,]| mi\\.", "", data$mileage))
```

Updated the name of milage column to -> mileage. Removed mi. and , to enable numeric operations

```r
empty_space_rows <-  data[rowSums(data == "", na.rm = TRUE) > 0, ]
cat("Empty Values:", nrow(empty_space_rows), "\n")
```

```
## Empty Values: 596
```

#summary statistics

```r
summary(data)
```

```
##     brand              model             model_year      mileage
##  Length:4009        Length:4009        Min.   :1974   Min.   :   100
##  Class :character   Class :character   1st Qu.:2012   1st Qu.: 23044
##  Mode  :character   Mode  :character   Median :2017   Median : 52775
##                                        Mean   :2016   Mean   : 64718
##                                        3rd Qu.:2020   3rd Qu.: 94100
##                                        Max.   :2024   Max.   :405000
##
##   fuel_type            engine          transmission         ext_col
##  Length:4009        Length:4009        Length:4009        Length:4009
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##    int_col             accident      clean_title            price
##  Length:4009        Min.   :0.0000   Length:4009        Min.   :   2000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.:  17200
##  Mode  :character   Median :0.0000   Mode  :character   Median :  31000
##                     Mean   :0.2459                      Mean   :  44553
##                     3rd Qu.:0.0000                      3rd Qu.:  49990
##                     Max.   :1.0000                      Max.   :2954083
##
##    horsepower      displacement    cylinders         engine_type
##  Min.   :  70.0   Min.   :0.650   Length:4009        Length:4009
##  1st Qu.: 248.0   1st Qu.:2.500   Class :character   Class :character
##  Median : 310.0   Median :3.500   Mode  :character   Mode  :character
##  Mean   : 332.3   Mean   :3.711
##  3rd Qu.: 400.0   3rd Qu.:4.700
##  Max.   :1020.0   Max.   :8.400
##  NA's   :810      NA's   :396
```

Unsupervised Learning

Prediction Models

Open-Ended Question/Conclusion