Predictive Modeling and Price Analysis of Used Cars: A Comprehensive Statistical Approach Using R

Ali Hamza

EC Utbildning – Data Scientist

R Programmering

2024/09

ABSTRACT

This research performs a comprehensive analysis of car price data from a dataset obtained through Blocket, focusing on cleaning, exploration, and statistical modeling. The primary objective is to investigate the relationships between vehicle characteristics such as mileage, horsepower, model year, drivetrain, and car prices. The analysis begins with data cleaning, including handling missing values, removing unwanted columns, and converting prices into numeric form. Visualizations as histograms and scatterplots are created to provide insights into price distributions and trends over time.

The script applies Pearson correlation to quantify the relationships between different variables and car prices, highlighting significant correlations. Further, linear and multiple linear regression models are fitted to predict car prices based on features like horsepower, model year, and mileage. Various diagnostic tools such as residuals analysis, Q-Q plots, and skewness tests, assess the normality of the variables and the accuracy of the models. Transformations are applied to address skewness in the data, enhancing model performance.

Potential issues such as non-linearity and non-normality will also be presented.

The final models are evaluated using metrics like root mean square error (RMSE), mean absolute percentage error (MAPE), and R-squared to quantify prediction accuracy. The analysis demonstrates the importance of key factors like horsepower and mileage in determining car prices and provides a framework for predictive modeling in the automotive market.

Keyword: comprehensive analysis, data cleaning, data exploration, statistical modeling, linear regression, multiple linear regression, residuals analysis, Q-Q plots, skewness tests, non-linearity, non-normality, predictive modeling.

INNEHÅLLSFÖRTECKNING

INLEDNING

I dagens snabbt växande begagnatmarknad för bilar är korrekt prissättning avgörande för både säljare och köpare. Priser på begagnade bilar påverkas av en mängd faktorer som bilens märke, modellår, miltal, hästkrafter och skick. Att kunna förutsäga ett rättvist och korrekt pris baserat på dessa faktorer är viktigt för att säkerställa en konkurrenskraftig marknad.

Denna rapport syftar till att analysera och förutsäga priser för begagnade bilar genom att tillämpa statistiska metoder och prediktiv modellering med hjälp av R-programmering. Genom att använda ett dataset som innehåller bilspecifikationer från en plattform som Blocket, utforskas samband mellan bilens attribut och dess pris. Genom användning av olika statistiska verktyg, såsom korrelationsanalys och linjära regressioner, kommer vi att analysera hur variabler som hästkrafter, modellår och miltal påverkar priset. Målet är att utveckla en robust prediktiv modell som kan användas för att förutsäga priset på begagnade bilar med hög noggrannhet.

Denna studie inkluderar flera steg såsom datainsamling, datastädning, utforskande dataanalys, modellering och utvärdering av prediktionsmodeller. Resultaten från denna analys ger värdefulla insikter för aktörer på bilmarknaden och kan potentiellt användas för att optimera prissättningsstrategier.

DATAINSAMLING

För att analysera och modellera priser på begagnade bilar använde vi ett dataset som innehåller specifikationer och egenskaper för bilar som fanns till salu på Blocket. Detta dataset innehåller variabler som pris, märke, modellår, hästkrafter, miltal och flera andra faktorer som kan påverka bilens pris.

DATAIMPORT

Data för denna studie hämtades från en Excel-fil som innehåller information om olika bilannonser från Blocket. Filen med namnet *Blocket_cars.xlsx* laddades upp lokalt och importerades in i R för vidare analys.

Importen av data utfördes med hjälp av funktionen read_excel från biblioteket readxl, vilket gjorde det möjligt att läsa in data från Excel-format direkt till R arbetsmiljö. Här är koden som användes för att läsa in data:

```
# IMPORT DATA
#-----------------------------------------------------------------------------------------------------

file_path <- "C:/Users/ali_h/Desktop/R Programmering Ali Hamza/R Programmering/Blocket_cars.xlsx"
car_data <- read_excel(file_path)
```

Efter att datan importerats, inspekterades den med hjälp av funktioner som View() för att säkerställa att den laddades korrekt och för att få en första överblick över datasetets struktur och innehåll. Detta steg var avgörande för att förstå vilka kolumner som var relevanta för vidare analys, samt för att identifiera eventuella saknade eller oönskade värden som behövde hanteras i nästa steg av projektet.

SYFTE

Syftet med denna studie är att undersöka hur olika faktorer påverkar priset på begagnade bilar, med målsättningen att utveckla en statistisk modell som kan förutse bilens pris baserat på specifika egenskaper. Studien fokuserar på att analysera hur variabler såsom miltal, hästkrafter, modellår, och andra tekniska och fysiska parametrar korrelerar med priset.

METOD

Analysen genomfördes med hjälp av R och statistiska metoder för att bearbeta, visualisera och modellera data.

4.1 DATAFÖRBEREDELSE

För att säkerställa datakvalitet genomfördes en datastädning. Saknade värden identifierades och irrelevanta kolumner togs bort. Dessutom omformaterades textdata till gemener, och variabler som "Pris" konverterades till numeriska värden:

```
car_data_fixed$Pris <- as.numeric(gsub("[^0-9.]", "", as.character(car_data_fixed$Pris)))
```

4.2 MODELLTRÄNING & UTVÄRDERING

Modellen för att analysera och förutse bilpriser byggdes med hjälp av linjär regression.

4.2.1 LINJÄR REGRESSION

Linjär regression används för att analysera relationen mellan en beroende variabel (i detta fall pris på bilar) och en eller flera oberoende variabler (som hästkrafter, miltal och modellår). Målet är att hitta en linje som bäst beskriver hur den beroende variabeln förändras i relation till de oberoende variablerna.

4.2.2 RESIDUALANALYS

För att utvärdera modellens prestanda genomfördes residualanalys. Residualerna visualiserades och testades för normalitet genom histogram och QQ-plots.

4.3 PRESTANDAMÅTT

Modellens prestanda mättes med hjälp av följande metoder:

1. RMSE (Root Mean Square Error)
2. MAPE (Mean Absolute Percentage Error)

Resultaten från dessa prestandamått ger en indikation på hur väl modellen förutsäger priser.

RESULTAT

5.1 LINJÄR REGRESSION

Analysen visade att det fanns ett starkt positivt samband mellan hästkrafter och pris, medan miltal hade ett negativt samband med pris. Modellåret hade också en viss inverkan på priset, där nyare bilar hade högre priser.

5.2 RESIDUALANALYS

Residualanalysen visade att modellen fungerade väl, även om vissa residualer indikerade att det fanns variationer som modellen inte fångade fullt ut.

DISKUSSION

Studien visar att det finns tydliga samband mellan vissa egenskaper och priset på begagnade bilar. Det största sambandet observerades mellan hästkrafter och pris, där kraftigare motorer var förknippade med högre priser. Samtidigt minskade priset i takt med att bilens miltal ökade. Däremot kan modellen förbättras genom att inkludera fler variabler eller genom att använda mer avancerade metoder för att hantera icke-linjära samband.

SLUTSATS

Denna studie har visat att linjär regression är en effektiv metod för att analysera priset på begagnade bilar baserat på deras specifikationer. Genom att använda variabler som hästkrafter, miltal och modellår kan vi relativt precist förutse priset på en bil. Dock skulle modellen kunna förbättras genom att inkludera fler faktorer, som exempelvis bilens skick eller utrustningsnivå.

KÄLLOR

1. Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow : concepts, tools, and techniques to build intelligent systems (Second edition). O'Reilly.
2. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An Introduction to Statistical Learning: with Applications in R* (2nd ed.). Springer.
3. Wickham, H., & Grolemund, G. (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. O'Reilly Media.