Blocket used car market

Statistical regressionanalysis of used cars on the Swedish online marketplace, blocket.se

Marcus Eklund
EC Utbildning
Course exam- R Programming
202404

Abstract

Work in progress

Abbreviations and terms

Index

1 Introduction

Between 2012 and 2022 the amount of personal cars in Sweden has increased from 4 447 165 to 4 980 543 however since 2021 it has started to decrease. From 2021 to 2022 it decreased with 0.12% (Appendix B) which isn't much but if we look at newly registered cars it shows something more drastic. Between December 2019 and December 2020 newly registered cars decreased with almost 28% (Statistikmyndigheten 2021) and that trend has continued. March 2024 saw a similar decrease of 21% compared to the same period the previous year (Trafikanalys 2024) and this is predicted to continue throughout 2024 (Levin 2023). So cars overall are decreasing recently but only by a relatively small amount however newly registered cars are decreasing much more quickly and that might lead us to ask "how does the used car market look?".

The main purpose of this report is to analyze a part of the used car market in Sweden and make a predictive model for the asking price on one of Sweden's biggest online market places for used cars, blocket.se. To achieve this the following list of tasks and questions are relevant,

1. Collect data on the used car market from Blocket.
2. Create a predictive model for asking price.
3. Determine which variables/features are most significant for determining price.
4. Are there other potential variables outside of Blocket that could have a potential effect on price?

2 Theory

2.1 Statistical Learning

Statistical learning refers to tools for understanding data and can be split into two groups, supervised and unsupervised learning. In supervised learning, the algorithm learns from labeled data, meaning the data is already tagged with the correct answer. It learns to map the input data to the correct output label based on examples of input-output pairs. Supervised learning include both predicting categories (classification) and predicting continuous values (regression). In unsupervised learning on the other hand we don't know the desired output so the algorithm learns on unlabeled data and tries to find hidden patterns and connections without explicit guidance. (James et al., 2023)

2.2 Linear Regression

Within supervised learning for continuous values the simplest method is a linear regression. It is used to understand the relationship between a dependent variable and one or more independent variables. It assumes that the relationship is linear, meaning that changes in the independent variable(s) gives a proportional change in the dependent variable, and fits a straight line on the observed data points. This makes it a powerful tool for predicting future outcomes and understanding relationships between variables. If you only have a single independent variable it is called a simple linear regression and if you have multiple independent variables it is a multiple linear regression. (James et al., 2023)

2.2.1 Multiple Linear Regression

2.2.1.1 Potential problems

2.3 Evaluating Models

2.3.1 Train- test data

2.3.2 Evaluation metrics

2.4 Feature Selection

3 Method

Below are the steps of how this report came to be. For more details of the work I refer to the code in Appendix A or the accompanying .R file.

3.1 Data Collection and Exploration

The collection of data was done by a web-scraper built by a colleague. The following are the parameters for which cars should be included,

1. Cars made from 2000 and forward.
2. Selling price between 20 000 kr and 500 000 kr.
3. Only private sellers.
4. No work vehicles.

Once we had the parameters we decided on 15 variables that should be collected from each ad,

1. Id, ad id to be able to go back and look at the actual ad.
2. Brand, car brand (labeled Märke in the code).
3. Model, model of car (labeled as Modell in the code).
4. Fuel, type of fuel (labeled Bränsle in the code).
5. Gearbox, type of gearbox (labeled Växellåda in the code).
6. Mileage, how far the car had driven in Swedish miles, 1 Swedish mile equals 10 kilometers (labeled Miltal in the code).
7. Model year, the year of the model (labeled Modellår in the code).
8. Car type, type of car (labeled Biltyp in the code).
9. Drivetrain, if the car has 2 wheel drive or 4 wheel drive (labeled as Drivning in the code).
10. Horsepower, power of the engine (labeled HK in the code).
11. Color, color of the car (labeled Färg in the code).
12. Engine size, size of the engine (labeled Motorstorlek in the code).
13. Date in traffic, date when the car was first legally registered to be in traffic (labeled Datum.i.trafik in the code).
14. Region, region of Sweden the car is located (labeled Region in the code).
15. Price, asking price of the car (labeled Pris in the code).

The scraper collected 10 083 ads with 15 variables. An initial inspection was done of the data to see that it had collected everything correctly.

3.2 Cleaning and Transforming Data

To turn the data into something that could be modeled the following steps were taken,

1. Formatted all columns to their correct formats.
2. Removed all rows with missing values.
3. Created 2 new columns, Age and Days in traffic, based on Model year and Date in traffic respectively.
4. Removed columns Model, Model year, Engine size and Date in traffic.
5. Grouped all car brands with less than 50 observations into Other.

After these steps there were 9 449 ads left with 13 variables.

3.3 Model Creation and Evaluation
To predict the price of the cars the following steps were taken,

1. Split data into training (70%) and test (30%).
2. Removed Id column.
3. Created an initial model with all variables.
4. Inspected the model for potential problems.
5. Ran forward subset selection
6. Created a second model with 7 of the most significant variables.
7. Compared the 2 models.
8. Performed cross-validation on the best model.
9. Deployed the model on the test data.
10. Compared RMSE of the cross-validated training data against the test data.

3.4 Public Data Collection
For the stats in the intro, data was collected from Statistics Sweden (Statistikmyndigheten SCB) using their API (see Appendix B).

4 Results and discussion

5 Conclusions

Appendix A

Appendix B

```
## Collecting data from SCB ------------------------------------
d <- pxweb_interactive("https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/
                          TK1001/TK1001A/PersBilarA")

# PXWEB query
pxweb_query_list <-
    list("Region" = c("00"),
        "Agarkategori" = c("000"),
        "ContentsCode" = c("TK1001AB"),
        "Tid" = c("2012", "2013", "2014", "2015", "2016", "2017", "2018", "2019",
                   "2020", "2021", "2022"))

# Download data
px_data <-
    pxweb_get(url =
      "https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/TK1001/TK1001A/
      PersBilarA",
                query = pxweb_query_list)

# Convert to data.frame
px_data_frame <- as.data.frame(px_data, column.name.type = "text",
      variable.value.type = "text")

# Get pxweb data comments
px_data_comments <- pxweb_data_comments(px_data)
px_data_comments_df <- as.data.frame(px_data_comments)

# Cite the data as
pxweb_cite(px_data)

View(px_data_frame)
names(px_data_frame)
str(px_data_frame)

scb_df <- px_data_frame |>
    mutate(across(c("år", "Personbilar i trafik"), as.integer)) |>
    rename(personbilar_i_trafik = `Personbilar i trafik`) |>
    na.omit()

str(scb_df)
View(scb_df)

ggplot(scb_df, aes(år, personbilar_i_trafik)) +
    geom_line() +
    scale_y_continuous()

tail(scb_df$personbilar_i_trafik, n = 1) -   scb_df$personbilar_i_trafik[1]

diff_21_22 <- max(scb_df$personbilar_i_trafik) -
      tail(scb_df$personbilar_i_trafik, n = 1)
prcnt <- diff_21_22 / max(scb_df$personbilar_i_trafik)
prcnt
```

References

James G., Witten D., Hastie T. & Tibshirani R. (2023). *An Introduction to Statistical Learning with Applications in R* (2nd ed.). Springer Texts in Statistics.

Levin M. (2023). *Antalet nyregistrerade fordon fortsätter att minska de kommande åren* https://www.trafa.se/vagtrafik/korttidsprognoser-for-fordonsflottan-13848/

Statistikmyndigheten (2021). *Antalet nyregistrerade bilar minskade med knappt 28 procent i december* https://www.scb.se/hitta-statistik/statistik-efter-amne/transporter-och-kommunikationer/vagtrafik/fordon/pong/statistiknyhet/nyregistreringar3/

Tafikanalys (2024). *Stor nedgång för nyregistrerade personbilar i mars* https://www.trafa.se/om-oss/press/#/pressreleases/stor-nedgaang-foer-nyregistrerade-personbilar-i-mars-3314188