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

Abbreviations and terms

Index

1 Introduction

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 blocket.se. To achieve this the following list of tasks and questions are relevant,

1. Create a predictive model for asking price of used car adds on blocket.
2. Determine which variables/features are most significant for determining price.
3. Can we draw any previously unseen conclusions from the data?
4. ???

1.1 Underrubrik – Exempel

Text är skriven på formatet Calibri med textstorlek 11.

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

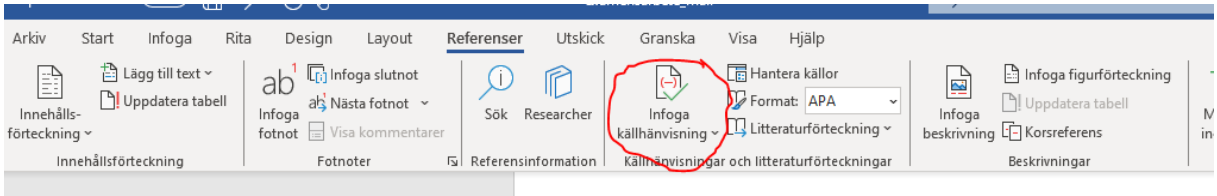
To predict the price of the cars the following steps were taken,

1. Splitting data into training (70%) and test (30%).
2. Creating an initial model

4 Results and discussion

RMSE för olika modeller	
Enkel Linjär Regression	xx
Lasso	xx
Ridge	xx

Tabell 1: Root Mean Squared Error (RMSE) för de fyra valda modellerna.



Figur 1: Hur man lägger in tabell eller figur nummer samt beskrivning.

5 Conclusions

Här besvarar du bl.a. frågeställningarna.

6 Teoretiska frågor

1. Kolla på följande video: https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s, beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.
Det är en graf som jämför fördelningen av vår data mot en normalfördelning. När punkterna ligger efter en diagonal linje så är vår data normalfördelad men om punkterna avviker från linjen så kan vi behöva transformera datan då våra modeller förväntar sig att datan den tränas på är normalfördelad.

2. Din kollega Karin frågar dig följande: ”Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?” Vad svarar du Karin?
Ja det stämmer. Maskininlärning handlar om att bygga prediktiva modeller som ger oss svar när vi stoppar in ny data. Statistisk regressionsanalys handlar om att skatta relationer mellan en beroende variabel och en eller fler oberoende variabler och med det skapa prediktiva modeller och dra slutsatser. Exempelvis så kan vi undersöka sambandet mellan ålder och lön och dra slutsatser om hur starkt sambandet är.

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?
Konfidenseintervall handlar om snittet för en viss grupp medans prediktionsintervall handlar om en individ och inkluderar osäkerheten I detta. Det betyder att prediktionsintervallet är alltid större än konfidensintervallet.

4. Den multipla linjära regressionsmodellen kan skrivas som:
$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_p x_p + \epsilon$
Hur tolkas beta parametrarna?
β_p för p≥1 tolkas som: ”Vad är effekten på Y när x_p ökar med en enhet, givet att alla andra variabler är fixa.”

5. Din kollega Hassan frågar dig följande: ”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du Hassan?
Ja BIC är ett sätt att skatta testfeltet genom att lägga till ett straff på träningsfelet för att kompensera för den bias som uppstår.

6. Förklara algoritmen nedan för ”Best subset selection”

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using using the prediction error on a validation set, C_p (AIC), BIC, or adjusted R^2. Or use the cross-validation method.

Best subset selection skapar en modell för varje enskild oberoende variabel och kombinationer av oberoende variabler och jämför dem med varandra för att hitta den bästa modellen. Bästa innebär den modell med lägst fel.

7. Ett citat från statistikern George Box är: "All models are wrong, some are useful." Förklara vad som menas med det citatet.
Ingen modell kan fånga all komplexitet I verkliga världen men dem kan ändå vara användbara. Alltså vi behöver inte förklara alla aspekter för att kunna ge ett bra estimat av verkligheten.

7 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.
2. Vilket betyg du anser att du skall ha och varför.
3. Något du vill lyfta fram till Antonio?

Appendix A

```
# load data
file_path <-
"C:/Users/marcu/WinCode/ec_utbildning/R/kunskapskontroll_r/car_
ads_data_02.csv"
raw_car_data <- read.csv(file_path)

# quick check on the data
View(raw_car_data)
summary(raw_car_data)
str(raw_car_data)
sum(is.na(raw_car_data))
dim(raw_car_data)
```

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.