R programmering med tillämpningar inom dataanalys

Kunskapskontroll

Amir Anissian

Detta individuell- och grupparbete var vår kunskapskontroll i kursen: R programmering med tillämpningar inom dataanalys. I vår Data Science utbildning på EC-utbildning,

Innehållsförteckning

Del 1

Introduktion

Detta grupparbete var en del av vår kunskapskontroll i kursen: R programmering med tillämpningar inom dataanalys. I vår Data Science utbildning på EC-utbildning, yrkeshögskola. Kursdatum: 2023-04-17 - 2023-05-26. Där en andra del bestod av en individuell modellering och rapport.

Grupparbetet gick ut på att samla information om bilars priser och egenskaper från Blocket.se och säkerställa att modellering av datan är möjligt genom en så kallad POC. Med hjälp av programmering i R.

Gruppmedlemmar: Amir Anissian, Natalia Makarova, Márk Mészáros, Victor F. Popa, Anders Pettersson och Tommy Nielsen.

Datainsamling

1.1 Syftet med modellen och vilken data behövs för det

Vårt mål är att bygga en prediktionsmodell som kan prediktera bilpriser för familjebilar (sedan, halvkombi, kombi) baserad på data som samlas in från Blocket.se, för årsmodeller mellan 2000–2011, sålda av privatpersoner.

Med limiterade resurser (tid) kommer vi att koncentrera oss i detta steg av projektet, sikta in oss på att med få grundläggande variabler. Försöka visa att det är möjligt att skapa en rimlig modell (POC) och att den kan förbättras genom modellering (individuell del av kunskapskontrollen). Tanken är att efter den slutgiltiga modelleringen så skulle man lägga ytterligare resurser på större och bredare datainsamling.

1.2 Vilken typ av data skall samlas in

Initialt tittade vi lite på web scraping som en potentiell metod för insamling. Det visar sig att vara något problematiskt för just Blocket.se. Detta samt risken att strida mot Blockets användarvillkor. Valde vi att påbörja manuell insamling av data.

Vi beslutade att samla in följande information kring bilarnas egenskaper:

- Bränsle: kategoriska data (Bensin/Diesel)

- Växellåda: kategoriska data (Manuell/Automat)

- Miltal: kategoriska data (intervall 0-10k 500mil, 10k-20k 1000 mil, 20-25k 5000 mil)

- Modellår: numerisk data (årtal → diskret)

- Biltyp: kategoriska data (sedan, halvkombi, kombi)

- Pris: numerisk data (i kronor → diskret)

Några andra egenskaper som vi diskuterade var:

Modell / märke: skulle snabbt innebära att vi får många variabler, vilket både skulle ge en väldigt komplicerad modellen och öka behovet av antal observationer (bilar). För att säkerställa tillräckligt med data för varje kombination av modell.

Prestanda variabler så som, hästkrafter, motorvolym. Då målgruppen är en familjebil där vi antar att prestanda inte är en prioritet. Samt att effekten av prestanda inte är lika stor för äldre bilar (2000 - 2011).

1.3 Hur mycket data skall samlas in

Baserat på övningar och projekt i andra kurser, samt bland annat erfarenheten från övningsexempel kursboken ”ISLR” i där datasetet AUTO (med 392 observationer). Kändes det rimligt att samla in ca 100 observationer per variabel för att säkerställa tillräckligt med data. Då vi valt ut 5 oberoende variabler och 1 responsvariabel, det kändes det rimligt med minimum 500 observationer i förhållande till variabler.

Målet blev därför att varje gruppmedlem skulle samla information från ca 100 bilar var. Vilket rimligtvis skulle göra att vi enkelt kunde nå målet (ca 500), även om problem skulle uppstå.

1.4 Kontrollera att datan vi samlat in är “rimlig”

Vid diskussion kring rimligheten. Kom vi fram till att modellen egentligen kan inte prediktera bilens slutpris. Eftersom man inte kan utgå ifrån vilket pris som säljaren och köparen kommer överens om. Ej heller om bilen ens blivit såld. Detta medför att vi inte kommer kunna prediktera ett eventuellt slutpris. Utan modellen kommer kunna prediktera ett rimligt utgångspris.

Utöver att i kommande EDA identifiera och bestämma ödet för orimlig data.

Fann vi att under insamlingen fick man en ”känsla” för vad som kunde tänkas vara orimlig data. Tex, utställningsbilar, reservdelsbilar. Vilket vi i efterhand diskuterat att en bra idé är att när man tillsammans påbörjar insamlingen. För att få möjlighet att hitta i alla fall de mest ”uppenbara” av dessa. För att säkerställa att eventuella skillnader i ex domänkunskap mellan personerna som samlar in datan nämnvärt påverkar insamlingen.

1.5 Samla in datan på ett konsistent sätt i gruppen

Variablerna och format för varje variabel definierades.
För att minimera ’human-error’ från säljaren eller data insamlare. Beslutade att inte ta med bilar där information har varit uppenbart tveksam eller saknats i någon av variablerna.

Däremot tog vi inte hänsyn till om säljaren felaktigt angett tex en VW Golf som ”småbil” och därför inte kom med eller en VW Polo (mindre modell) som ”halvkombi” så den kom med. Eftersom vi antar att detta kommer spegla både de privata säljarnas okunskap / beteende och hur Blockets struktur påverkar kvaliteten av datan. Detta medför en förenkling av datainsamlingen. Men att modellen inte kan förväntas prestera lika när det är företag som säljer på Blocket eller på en professionell bilsäljarsite.

För att minimera risken att samma bil registreras av olika personer. Delade vi in insamlingen så att varje person samlade in bilar från olika årsmodeller.

1.6 Hur skall vi organisera oss som grupp

Vi har diskuterat grundfrågorna och strategin tillsammans i hela gruppen.

Möten och delning av filer skulle ske genom Teams. Ett naturligt val då vi alla är bekväma med detta i vår utbildning och arbetet skedde hemifrån.

Datainsamlingen delades upp jämnt mellan gruppmedlemmarna.

1.7 Sammanfattning datainsamling

Format: Excel, innehållande 704 rader och 6 kolumner.

Fil: BlocketBilData [ii]

Variabel	Innehåll
Bränsle	"Bensin", "Diesel", "Miljöbränsle/Hybrid"
Växellåda	"Manuell", "Automat"
Miltal	“3000 – 49999” olika intervall och “Mer än 50 000”
Modellår	“Between 2000 - 2011”
Biltyp	"Sedan","Halvkombi","Kombi"
Pris	“2000 – 369000 Kr – Kr = valuta SEK.”

Proof Of Concept (POC)

Vi började med att kontrollera att datan, och att variablerna bara innehöll de värde som vi förväntade oss. Några fåtal felaktiga observationer fanns och togs därmed bort. Eventuella nollvärde togs också bort.

Variabeln miltal innehöll intervaller och vi bestämde att det vore rimligare att använda sig av mitten av intervallet istället, till exempel 0–499, 500 – 999..... Eftersom miltal i verkligheten är en enskild siffra inom samma intervall och att vi antar att förändringen av miltal inom ett (samma) intervall inte utgör någon större skillnad i effekt på priset. Vi väljer medelvärdet (mitten) för att uppskatta miltalet för varje observation.

Datasetet delades upp i train- och test-set. För att kunna få en rättvis bedömning av den slutgiltiga modelns (efter del 2) prestanda. Stratifierades train/test uppdelninen baserat på ”Modellår”. För att säkerställa att både train och test datan är representativ (innehåller bilar från varje årsmodell).

En (o-regulariserad) linjär regressionsmodell tränades och utvärderades med cross validation.

Med hjälp av hypotesprövningen i funktionen ”summary” på modellen. Ser vi en indikation på att minst några av variablerna och modellen i sig har en betydande effekt på priset.

För att kunna utvärdera prestandan på modellen använder vi oss av RMSE. RMSE använder vi då det ger en lättolkad indikation för inom vilket intervall som vi förväntar att modellens predikterade värde kommer att skilja sig från verkligheten.

Resultat (cross-validation): 25293

Slutsats: Vi förväntar oss att prediktionerna som modellen gör, normalt inte kommer att skilja sig mer än + / - 25293 kr.
Med fortsatt modellering (del 2 av kunskapskontrollen) hoppas vi att kunna förbättra detta något. Men eftersom vi har ett fåtal variabler som täcker ett stort antal olika egenskaper. Förväntar vi oss att det kan vara svårt att få fram en extremt bra modell. Däremot kan vi förvänta att modellen generaliserar bra över ett större spektrum av bilar.

R kod för POC

Kod för POC i R finns tillgänglig [iii]

Del 2 – Regressionsanalys

Individuell del

Introduktion

- Bakgrund

Jag vill använda data som vi har samlat in för att göra regressionsanalys.
Regressionsmodellen vi använder är baserad på linjär regression för att förutsäga bilpriser. Modellen verkar innehålla flera variabler som kan påverka bilens pris, inklusive bränsletyp, årsmodell och miltal.

- Syfte och Frågeställning

Jag skulle vilja sälja kunden en lösning som erbjuder rimliga priser för att sälja eller köpa bilar.

Databeskrivning / EDA (Exploratory Data Analysis)

Som vi presenterade i del 1 använde vi ett dataset som vi samlat in som grupp på blocket.se. Variablerna vi samlar in är: "pris", "bränsle", "växellåda", "miltal", "modellår", "biltyp". Jag har inga dubbletter eller nollvärden, men jag har några outliers. "miltal" var som intervall så konverterade jag som genomsnittet. Dessutom har jag några andra kategoriska variabler som jag konverterade dem till numeriska värden med hjälp av one-hot encoding. Dessa variabler efter överföring är "Miltal_mitten", "bränsle", "växellåda" och "biltyp".

Här kan ni se fem rader av vår dataset.

```
head(data)
# A tibble: 6 × 6
  Bränsle Växellåda Modellår Biltyp  Pris Miltal_mitten
  <chr>   <chr>        <dbl> <chr>  <dbl>         <dbl>
1 Bensin  Manuell        2006 Sedan  59500         20250
2 Bensin  Automat        2006 Sedan  40000         17250
3 Bensin  Manuell        2007 Sedan  29900         13750
4 Bensin  Manuell        2006 Sedan  85000          9250
5 Bensin  Manuell        2007 Sedan  33800         21250
6 Bensin  Automat        2007 Sedan  25000         21750
```

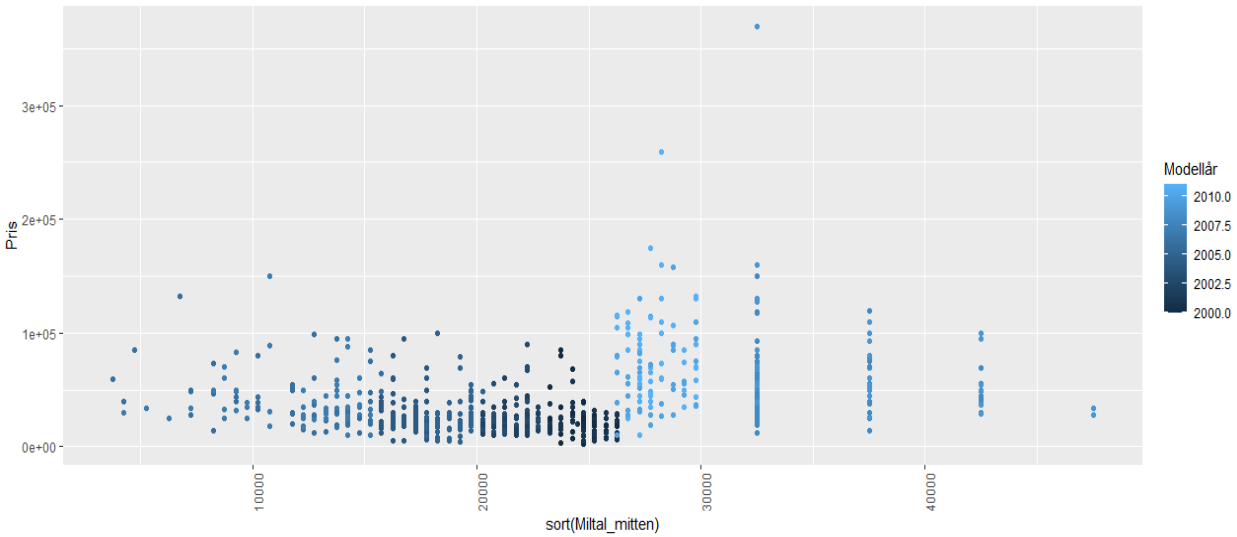
Vår dataset har 698 observation och 6 kolumnen.

```
> dim(data)
[1] 698   6
```

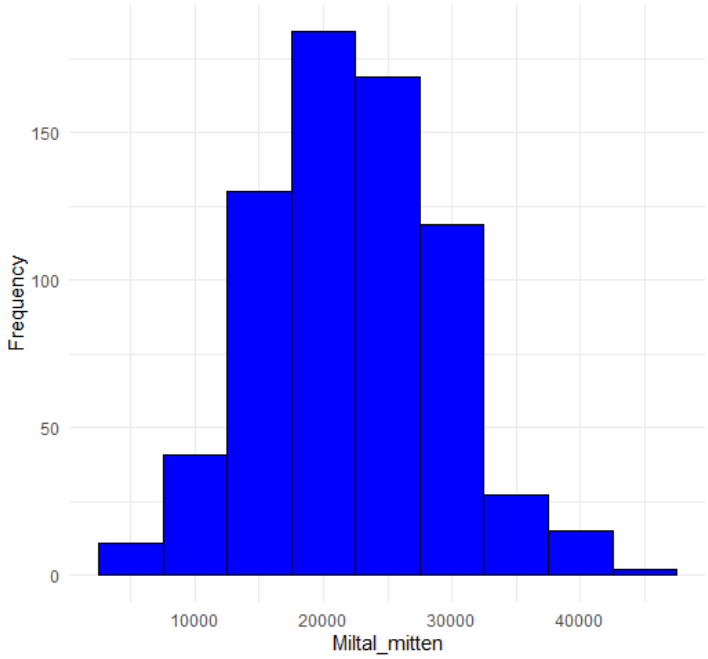
Denna bild visar sammanfattning av vår numeriska data

```
> summary(data)
 Bränsle.BränsleBensin  Bränsle.BränsleDiesel Växellåda.VäxellådaAutomat  Växellåda.VäxellådaManuell    Modellår
 Min.   :0.000000       Min.   :0.000000      Min.   :0.0000000          Min.   :0.0000000          Min.   :2000
 1st Qu.:0.000000       1st Qu.:0.000000      1st Qu.:0.0000000          1st Qu.:0.0000000          1st Qu.:2003
 Median :1.000000       Median :0.000000      Median :0.0000000          Median :1.0000000          Median :2005
 Mean   :0.739255       Mean   :0.260745      Mean   :0.3037249          Mean   :0.6962751          Mean   :2005
 3rd Qu.:1.000000       3rd Qu.:1.000000      3rd Qu.:1.0000000          3rd Qu.:1.0000000          3rd Qu.:2008
 Max.   :1.000000       Max.   :1.000000      Max.   :1.0000000          Max.   :1.0000000          Max.   :2011
 Biltyp.BiltypHalvkombi  Biltyp.BiltypKombi   Biltyp.BiltypSedan       Pris         Miltal_mitten
 Min.   :0.000000        Min.   :0.0000000    Min.   :0.0000000      Min.   :  2000   Min.   : 3750
 1st Qu.:0.000000        1st Qu.:0.0000000    1st Qu.:0.0000000      1st Qu.: 19900   1st Qu.:17250
 Median :0.000000        Median :0.0000000    Median :0.0000000      Median : 29800   Median :22250
 Mean   :0.269341        Mean   :0.4183381    Mean   :0.3123209      Mean   : 39529   Mean   :22617
 3rd Qu.:1.000000        3rd Qu.:1.0000000    3rd Qu.:1.0000000      3rd Qu.: 50000   3rd Qu.:27250
 Max.   :1.000000        Max.   :1.0000000    Max.   :1.0000000      Max.   :369000   Max.   :47500
```

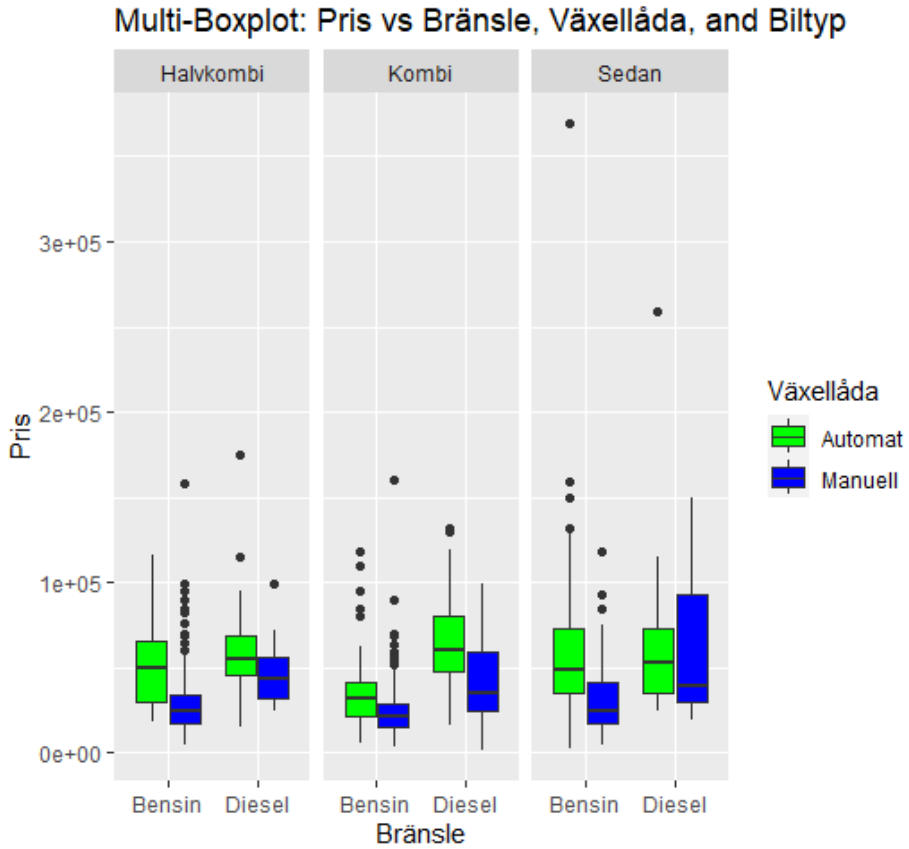
I figuren nedan kan ni se fördelning av bilpris baserad av Miltal och Modellår.



Här ser ni att miltal har en normalfördelningskurva som visar vår data fördelade normalt.



Denna bild nedan visar biltyp, Bränsle och Växellåda vs Pris.



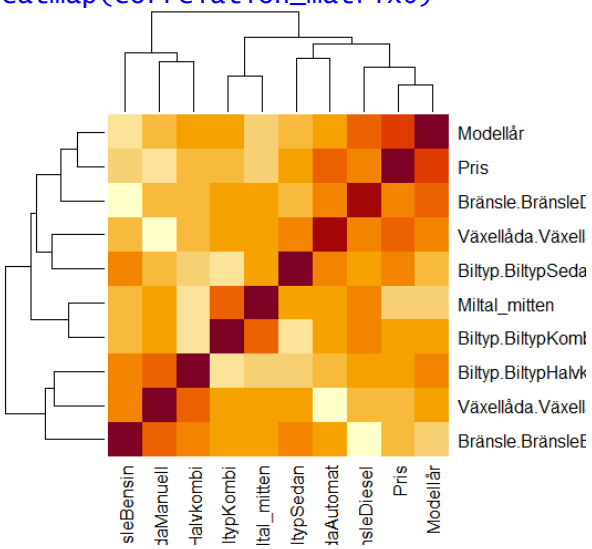
Generellt kan man se att dieselbilar med automatlåda är dyrare

Här kan man se korrelation matrix som visar att det finns mest korrelation mellan Pris och modellår.

```
> correlation_matrix0 <- cor(data)
                                 Pris
Bränsle.BränsleBensin      -0.28243130
Bränsle.BränsleDiesel       0.28243130
Växellåda.VäxellådaAutomat  0.35378293
Växellåda.VäxellådaManuell -0.35378293
Modellår                    0.55203606
Biltyp.BiltypHalvkombi     -0.06536208
Biltyp.BiltypKombi         -0.07728728
Biltyp.BiltypSedan          0.14483110
Pris                        1.00000000
Miltal_mitten              -0.27459151
```

Figuren visar en korrelation matrix heatmap och det finns mest korrelation mellan Pris och modellår .

```
> heatmap(correlation_matrix0)
```



Metod och Modeller (Teori)

- **One-hot encoding:** För att skapa ett binärt attribut per kategori: ett attribut som är lika med 1 och 0 annars, och så vidare. Detta kallas one-hot encoding, eftersom endast ett attribut kommer att vara lika med 1 (hot), medan de andra kommer att vara 0 (cold).[iv]
- **linear model,** *lm()*: används för att anpassa linjära modeller, inklusive multivariate sådana. Den kan användas för att utföra regression, enkel stratum-analys av varians samt analys av kovarians.[v]

  model_lm <- lm(Pris ~ ., data = data, subset = X_train_strat_index)

- **Fitting Generalized Linear Models,** *Glm()*: Används för att anpassa generaliserade linjära modeller genom att ange en symbolisk beskrivning av den linjära prediktorn och en beskrivning av felfördelningen. [vi]

  model_glm <- glm(Pris ~ ., data = data, subset = X_train_strat_index)

- **Cross Validation:** en teknik som används inom maskininlärning och statistisk modellering för att bedöma prestanda och generaliseringsförmåga hos en prediktiv modell. Den innebär att tillgängliga datamängder delas upp i flera delmängder, vanligtvis kallade "folds".

- **Lasso-regression:** som står för Least Absolute Shrinkage and Selection Operator Regression, är en regulariserad version av linjär regression. Den lägger till en regulariseringsterm till kostnadsfunktionen, men använder $\ell 1$-normen av viktvektorn i stället för hälften av kvadraten på $\ell 2$-normen.[vii]

  $$J(\mathbf{\theta}) = \text{MSE}(\mathbf{\theta}) + \alpha \Sigma_{i=1}^{n} |\theta_i|$$

- **Ridge:** Ridge-regression (även kallad Tikhonov-regulering) är en regulariserad version av linjär regression: en regulariseringsterm lika med α $\sum_{i=1}^{n}$ θi^2 läggs till kostnadsfunktionen. Detta tvingar in inlärningsalgoritmen att inte bara anpassa sig till datan, utan också hålla modellvikterna så små som möjligt.[viii]

  $$J(\mathbf{\theta}) = \text{MSE}(\mathbf{\theta}) + \alpha \frac{1}{2} \Sigma_{i=1}^{n} \theta_i^2$$

Denna är kod som jag använde att utföra Lasso och Ridge. När vi sätter alpha=1 har vi Lasso modellen och när vi sätter alpha=0 har vi Ridge modellen

```
lasso_model <- cv.glmnet(scaled_train_strat, y_train_strat, alpha = 1)
ridge_model <- cv.glmnet(scaled_train_strat, y_train_strat, alpha = 0)
```

- RMSE: *Root Mean Squared Error* är roten hur genomsnitts felet på alla gissningar som modellen har gissat.

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

Projekt Resultat och Analys

I del 1 fick vi ett intressant resultat. Vi fick bättre resultat i glm än cv.glm. Jag tror att anledningen till detta är därför använde vi delta [1]. Det betyder att vi använde första modellen i stället för den bästa.

```
> rmse_data
# A tibble: 2 × 2
  Data_set   RMSE
  <chr>     <dbl>
1 Train    24838.
2 CV       25293.
```

De är mitt RMSE resultat:

```
lm RMSE: 24838.37
glm RMSE: 24838.37
Lasso RMSE: 23520.77
Ridge RMSE: 24053.05
```

- Både linjär regression (lm) och generaliserad linjär modell (glm) ger liknande resultat med en RMSE på 24 838.37. Det innebär att dessa modeller har en liknande förmåga att förutsäga priset baserat på de tillgängliga prediktorerna.
- Lasso-regressionen har en någon lägre RMSE på 23 520.77, vilket indikerar att den modellen har bättre prestanda jämfört med de andra linjära modellerna (lm och glm).
- Ridge-regressionen har en RMSE på 24 053.05, vilket är något högre än Lasso-regressionen men fortfarande bättre än de andra linjära modellerna (lm och glm).

Slutsats och förslag på potentiell vidareutveckling

Förslag på potentiell vidareutveckling:

- Inkludera fler relevanta variabler: Det kan vara fördelaktigt att undersöka om det finns ytterligare variabler som kan påverka bilpriset. Till exempel kan vi överväga att inkludera motorstorlek, tillverkarmärke eller geografisk plats.

- Hantera potentiella icke-linjära relationer: I vissa fall kan relationen mellan variablerna och bilpriset vara icke-linjär. Det kan vara användbart att undersöka möjligheten att lägga till icke-linjära termer eller utföra en icke-linjär regressionsanalys för att fånga sådana relationer.

Genom att genomföra dessa förslag kan vi förbättra modellens noggrannhet och relevans vid förutsägelse av bilpriser och få en djupare förståelse för de faktorerna som påverkar prissättningen på begagnade bilar.

Appendix / referens

[i] ISLR; ”An Introduction to Statistical Learning - with Applications in R”
ISBN 978-1-4614-7137-0, tillgänglig: https://www.springer.com/series/417

[ii] BlocketBilData
https://github.com/amiranissian/R/tree/main/Bilpris/data/BlocketBilData.xlsx

[iii] https://github.com/amiranissian/R/tree/main/Bilpris

[iv] Géron, Aurélien. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed. O'Reilly Media, 2019.

[v] RStudio,2023.03.0 Build 386
[vi] RStudio,2023.03.0 Build 386

[vii] Géron, Aurélien. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed. O'Reilly Media, 2019.

[viii] Géron, Aurélien. Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. 2nd ed. O'Reilly Media, 2019.