Contents

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.

Gruppmedlemar: 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 ett 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]

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.xlsx – separat fil.

[iii] R kod för POC – https://github.com/NataliaMak20/Blocket-project-with-R

Del 2.

1. Introduction.

Jag antar att det finns en kund som köper bilar av privatpersoner och sysslar med alla typer av fordon. Baserat på personlig erfarenhet kan det vara svårt att bestämma ett rimligt pris för en bil. Därför tror jag att det skulle vara användbart att ha tillgång till en modell som kan hjälpa till i detta avseende.

För att möta detta behov har jag samlat in data om kombi, halvkombi och sedaner listade på Blocket mellan 2000 och 2011. Med denna information siktar jag på att utveckla en lösning som kan ge kunden korrekta prognoser för rimliga bilpriser när det gäller att köpa, sälja eller byta fordon.

2. Databeskrivning / EDA (Exploratory Data Analysis)

Som grupp har vi samarbetat för att samla in ett dataset från webbplatsen blocket.se. Datauppsättningen innehåller olika variabler som pris, bränsletyp, växellåda, modell och biltyp. Vi har sett till att det inte finns några dubbletter av poster eller nollvärden, även om det finns några extremvärden.

För de fall där körsträckan sjönk mellan 10 000 och 10 500, beräknade vi medelvärdet. Dessutom, eftersom vi har kategoriska variabler i vår datauppsättning, såsom Miltal_mitten, bränsletyp, växellåda och biltyp, har vi omvandlat dem till numeriska värden med hjälp av one-hot-kodning.

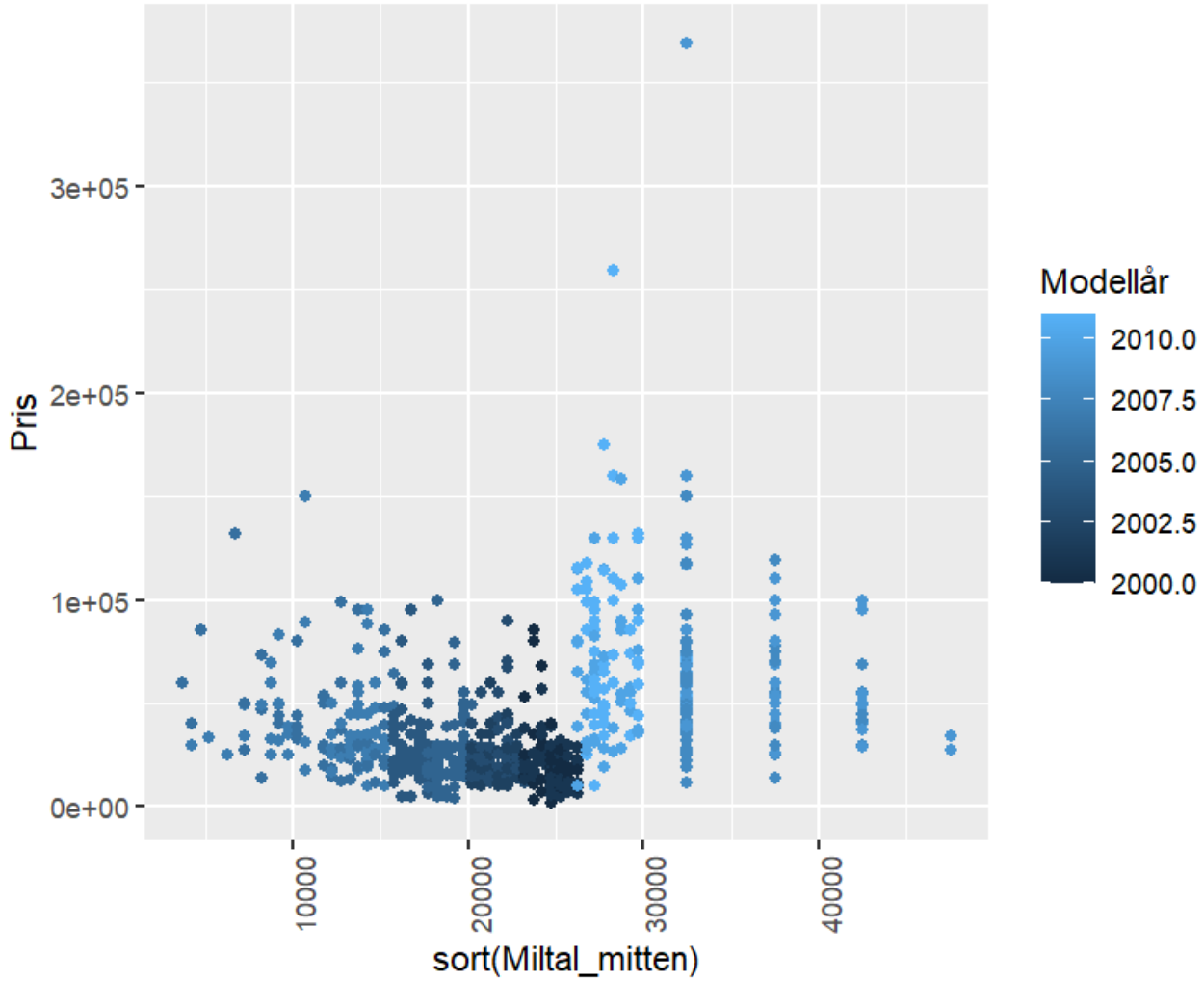
- I bilden nedan kan vi se de första 6 raderna av datamängden som vi har samlat in.

	Bränsle	Växellåda	Miltal	Modellår	Biltyp	Pris
	<chr>	<chr>	<chr>	<dbl>	<chr>	<dbl>
1	Bensin	Manuell	20 000 - 20 499	2006	Sedan	59500
2	Bensin	Automat	17 000 - 17 499	2006	Sedan	40000
3	Bensin	Manuell	13 500 - 13 999	2007	Sedan	29900
4	Bensin	Manuell	9 000 - 9 499	2006	Sedan	85000
5	Bensin	Manuell	21 000 - 21 499	2007	Sedan	33800
6	Bensin	Automat	21 500 - 21 999	2007	Sedan	25000

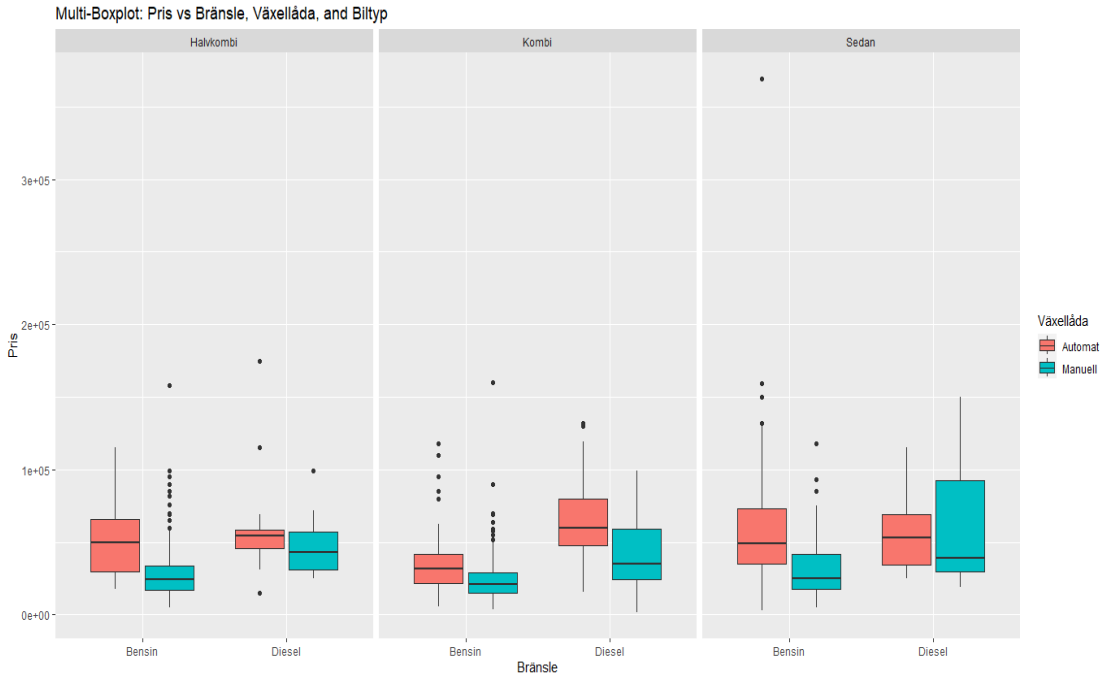
- Här ser vi dimensionen på datasetet, vilket är 704 observationer och 6 kolumner.

```
> dim(data)
[1] 704   6
```

- På bilden nedan kan vi se två otliers som vi tagit bort i efterhand.



- Den här bilden visar våra biltyper och vilken typ av bränsle som generellt är dyrare och vilken typ av växellåda som är dyrare



- Den här bilden visar kvartilerna i varje kolumn, minimivärdet, den första kvantilen är 25 %, medianen, medelvärdet, den tredje kvantilen är 75 % och maxvärdet. Den visar om kolumnen är kategorisk och sedan visar den bara vilken klass det är.

```
  Bränsle           Växellåda           Miltal
Length:704         Length:704         Length:704
Class :character   Class :character   Class :character
Mode  :character   Mode  :character   Mode  :character


   Modellår       Biltyp             Pris
Min.   :2000   Length:704         Min.   :  2000
1st Qu.:2003   Class :character   1st Qu.: 19800
Median :2005   Mode  :character   Median : 29400
Mean   :2005                      Mean   : 39659
3rd Qu.:2008                      3rd Qu.: 50000
Max.   :2011                      Max.   :369000
```

3. Metod och Modeller (Teori)

Linjär modell: Används för att anpassa linjära modeller, inklusive multivariate. Linjär modellering används för att utföra regression, analys av varians och kovarians.

Glm: Används för att anpassa generaliserade linjära modeller genom att ange en symbolisk beskrivning av den linjära prediktorn och en beskrivning av felfördelningen.

One-hot encoding: Skapa nya binära variabler för varje unik kategori i den ursprungliga variabeln. För varje kategori skapar vi en ny variabel som har värdet 1 om observationen

tillhör den kategorin och 0 annars. På så sätt representeras varje kategori som en separat variabel med enkelt numeriska värden.

Cross Validation: En metod som används för att utvärdera prestandan hos en modell och för att bedöma hur bra den generaliserar osett data. Istället för att bara dela in datamängden i en träningsdel och en testdel, delas den in i flera mindre delar, vanligtvis kallade "fold".

Lasso: En metod som hjälper till att välja och reducera antalet variabler i en modell genom att minska eller eliminera koefficienterna för mindre betydelsefulla variabler, vilket leder till en mer enkel och robust modell.

Ridge: En metod som begränsar koefficienternas storlek genom att lägga till en straff i modellen. Det hjälper till att reducera överanpassning genom att balansera betydelsen av olika variabler och skapa en mer robust och generaliserbar modell.

RMSE: *Root mean squared error* är roten hur genomsnitts felet på alla gissningar som modellen har gissat.
Formula: RMSE = sqrt(mean((predicted - actual)^2))

4. Resultat och Analys.

Lasso RMSE: 26987. Dvs att vi i genomsnitt gör en felbedömning på pris med 26.987 kr med hjälp av vår lasso modell.

Ridge RMSE: 28367,76. Dvs att vi i genomsnitt gör en felbedömning på pris med 28.367,76 kr med hjälp av vår ridge modell.

5. Slutsats och förslag på potentiell vidareutveckling.

Med tanke på omständigheterna vi stod inför tror jag att resultaten som genereras av vår topppresterande modell är logiska. I min situation visade sig lassomodellen vara den mest effektiva, med ett genomsnittligt prediktionsfel på cirka 27 000 kr. Med tanke på en bil prissatt till 200 000 kr kan ett prediktionsfel på 27 000 kr anses rimligt.

Det finns dock områden där förbättringar kunde ha gjorts. Det skulle ha varit fördelaktigt att samla en större urvalsstorlek, utforska ytterligare variabler och utveckla en mer komplicerad modell. Dessutom kunde testning av alternativa modeller som stödvektormaskin, gradientförstärkning och ensemblemetoder ha gett värdefulla insikter och potentiellt förbättrat resultaten.

1. **Utmaningar du haft under arbetet samt hur du hanterat dem.**
   Utmaningar för det här uppgift var mindre tid för det, eftersom jag tycker att uppgiften var ganska komplext. Också hur vi ska separera gruppdel från individuelt, och sedan kombinera det för slutlig kod och rapport. Men jag kunde diskutera allt det där med mina klasskamrater och fick bra hjälp av dem.

2. **Vem som varit i din grupp och en reflektion om hur grupparbetet fungerat.**
   Vår grupp var väldigt strukturerad och organiserad. Vi träffades nästan varje dag på Teams och diskuterat tillsammans allt vi gjort. Allt arbete delades lika mellan gruppmedlemmarna. Jag tror att tiden var den enda utmaningen för det här uppdraget, men eftersom det var många personer i gruppen gjorde vi allt snabbt och effektivt.

3. **Vilket betyg du anser att du skall ha och varför.**
   Baserat på den ansträngning och kunskap jag har investerat i det här uppgiften anser jag att jag förtjänar betyget "VG".

4. **Tips du hade ”gett till dig själv” i början av kursen nu när du slutfört den.**
   Jag skulle säga att det är viktigt att inte fastna på något, om det finns ett fel eller koden inte fungerar, utan fortsätt att studera, testa och utforska.