Project in R Programming Language

Multiple Regression Analysis to Predict Car Price

Student: Muhammad Mahmudur Rahman

Kurs: R programmering för dataanalys

Kurslärare: Antonio Prgomet

Utbildning: Data Scientist

Skola: EC Utbildning, Stockholm, Sweden

Datum: 2024-04-26

Abstract

The scope of the project is implementation of multiple regression analysis through R programming language. To do the analysis, secondhand car data was collected from www.blocket.se. It was collected data from eight different selected counties of Sweden, afterwards multiple regression analysis was applied to predict the car price. Three different models were created to predict the car price where Model 1 showed the best performance among the three models and predicted the most accurate result.

Innehållsförteckning

1 Inledning

Målet med projektet var förutsägelse av begagnade bilpriser genom multipel regressionsanalys. Totalt 700 data av begagnade bilar samlades in manuellt genom grupparbeten från åtta olika län i Sverige som är Stockholm, Dalarna, Östergötland, Skåne, Kalmar, Jönköping, Gotland och Västra Götaland från www.blocket.se där prisklassen var 100 000 kr till 500 000 kr. Funktionerna i datan var Car_Price, County, Fuel, Gearbox, Mileage, Model_Year, Car_Type, Drivetrain_System, Horsepower, Color, Brand and Model. Data samlades in från www.blocket.se, så den främsta intressanta delen av projektet var regressionsanalys med den nya datan där inga projekt gjorts tidigare. R är ett kraftfullt programmeringsverktyg som i hög grad kan hjälpa analytisk forskning på en mängd olika sätt. Regressionsanalys är en av tillämpningarna som tillämpades i detta projekt och förutspådde bilpriser med hög noggrannhet. Nedan är uppgifterna som samlades in från www.blocket.se:

	A	B	C	D	E	F	G	H	I	J	K	L
1	Car_Price	County	Fuel	Gearbox	Mileage	Model_Yea	Car_Type	Drivetrain_	Horsepow	Color	Brand	Model
2	419 000	Stockholm	Miljöbräns	Automat	14 452	2017	SUV	Fyrhjulsdri	412	Silver	Volvo	XC90
3	329 800	Stockholm	Diesel	Automat	11 818	2018	SUV	Fyrhjulsdri	192	Svart	Volvo	XC60
4	239 800	Stockholm	Miljöbräns	Automat	11 188	2016	SUV	Fyrhjulsdri	198	Mörkgrå	Toyota	RAV4
5	269 800	Stockholm	Bensin	Automat	3 748	2021	Halvkombi	Tvåhjulsdri	136	Svart	Mercedes-	A
6	239 800	Stockholm	Bensin	Automat	13 972	2017	Halvkombi	Fyrhjulsdri	252	Svart	BMW	330
7	174 800	Stockholm	Bensin	Manuell	4 561	2018	Halvkombi	Tvåhjulsdri	127	Vit	Honda	Civic
8	238 900	Stockholm	El	Automat	1 528	2022	Halvkombi	Tvåhjulsdri	150	Grå	Nissan	Leaf
9	149 800	Stockholm	Diesel	Manuell	8 961	2016	SUV	Tvåhjulsdri	111	Vit	Nissan	Qashqai
10	189 800	Stockholm	Diesel	Automat	14 780	2015	SUV	Fyrhjulsdri	184	Blå	Volvo	XC60

Fig 1. Insamlade data från Blockets hemsida

2 Teori

2.1 Multiple Regression Analysis[1]

Multipel regressionsanalys är en statistisk metod som används för att undersöka sambandet mellan en beroende variabel och två eller flera oberoende variabler.

I multipel regressionsanalys är målet att modellera förhållandet mellan den beroende variabeln (variabeln du försöker förutsäga) och flera oberoende variabler (variablerna som används för att förutsäga den beroende variabeln). Modellekvationen tar formen:

Y = β0 + β1X1 + β2X2 + … + βnXn + ϵ

Här:

- Y är den beroende variabeln.
- X 1 , X 2 , … , X n X 1 , X 2 ,..., X n är de oberoende variablerna.
- β 0 , β 1 , β 2 , … , β n β 0 ,β 1 , β 2 ,...,β n är regressionskoefficienterna som representerar sambandet mellan de oberoende och beroende variablerna.
- ϵ är feltermen.

2.2 R-squared (R^2)[2]

R-squared (R^2) är ett statistiskt mått som används för att bedöma passformen hos en regressionsmodell. Den representerar andelen av variansen i den beroende variabeln som förklaras av de oberoende variablerna i modellen. R-kvadratvärden sträcker sig från 0 till 1, där:

- 0 indikerar att modellen inte förklarar någon av variabiliteten för den beroende variabeln kring dess medelvärde.
- 1 indikerar att modellen förklarar all variation av den beroende variabeln kring dess medelvärde.

2.3 RMSE[2]

Root Mean Square Error (RMSE) är ett mått på skillnaderna mellan värden som förutsägs av en modell och de faktiska observerade värdena. Den representerar kvadratroten av de genomsnittliga kvadrerade skillnaderna mellan förutsagda och observerade värden. RMSE beräknas enligt följande:

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

Här:

• n är antalet observationer.

• yi är det faktiska observerade värdet.

• ŷ är det förutsagda värdet av modellen.

2.4 Bayesian Information Criterion (BIC)[2]

Bayesian Information Criterion (BIC) är ett kriterium för modellval bland en ändlig uppsättning modeller. Den balanserar modellens goda passform med modellens komplexitet. BIC straffar modeller för att ha fler parametrar och gynnar enklare modeller som fortfarande förklarar data väl. BIC beräknas enligt följande:

BIC=−2ln(L)+kln(n) där:

• L är sannolikheten för modellen givet data.

• k är antalet parametrar i modellen.

• n är provstorleken.

2.5 Confidence Interval (CI)[3]

Ett konfidensintervall tillhandahåller ett intervall av värden inom vilka den sanna populationsparametern (som medelvärde eller regressionskoefficient) sannolikt faller med en viss konfidensnivå.

I regressionsanalys indikerar ett konfidensintervall runt det förutsagda medelsvaret vid ett givet värde av den eller de oberoende variablerna det intervall inom vilket vi förväntar oss att det sanna medelsvaret faller med en viss konfidensnivå.

Till exempel betyder ett 95 % konfidensintervall för en regressionsförutsägelse att om vi skulle prova från populationen upprepade gånger och beräkna konfidensintervall för varje urval, skulle vi förvänta oss att 95 % av dessa intervall innehåller det sanna medelsvaret.

2.6 Prediction Interval (PI)[3]

Ett prediktionsintervall tillhandahåller en rad värden inom individuella framtida observationer sannolikt kommer att falla med en viss nivå av konfidens.

I regressionsanalys indikerar ett prediktionsintervall runt ett specifikt förutsagt värde för den beroende variabeln vid ett givet värde av den oberoende variabeln det intervall inom vilket vi förväntar oss att det faktiska observerade svaret för en ny observation ska falla med en viss nivå av konfidens.

Förutsägelseintervall är vanligtvis bredare än konfidensintervall eftersom de tar hänsyn till både variabiliteten hos data och osäkerheten vid skattning av modellparametrar, såväl som variabiliteten hos individuella framtida observationer.

Till exempel betyder ett 95 % prediktionsintervall för en regressionsprediktion att vi förväntar oss att 95 % av framtida observationer faller inom det intervallet.

2.7 API[4]

API, står för Application Programming Interface, är en uppsättning regler eller protokoll som gör det möjligt för mjukvaruapplikationer att kommunicera med varandra för att utbyta data, funktioner och funktionalitet.

3 Metod

3.1 Datainsamling

Grupp och gruppmedlemmar

Data för att förutsäga bilpriset med multipel regression har samlats in manuellt från webbplatsen www.blocket.se. Denna datainsamling var ett grupparbete. Vi var åtta gruppmedlemmar totalt inklusive mig i grupp nummer ett som arbetade med att samla in data. Förutom mig var mina gruppmedlemmar Arina Godman, Filip Östlund, Isabella Frid, Shangchanhui Feng, Shriya Walia, Turzo Khan och Wissam Rateb.

Grupparbete

Så fort gruppen skapades började vi arbeta tillsammans. Vi hade regelbunden kontakt genom Teams. Vi hade möte och delade upp uppgifterna mellan medlemmarna. Vi bestämde att varje gruppmedlem kommer att samla in 100 data var från www.blocket.se.

Fördelar med grupparbete

Diskussion - det var den bästa delen för mig att arbeta i gruppen. Olika professionell har olika idéer, skicklighet, kunskaper och erfarenheter eftersom vi har olika yrkesbakgrund. Genom grupparbetet är det möjligt att få andras kunskaper och idéer. Dessutom har grupparbete vissa förtjänster, olika perspektiv, ökad kreativitet och förbättrad kommunikationsförmåga.

Styrkor och utveckling att arbeta i grupp

• Analysera och bearbeta snabbt stora mängder information som kan vara till hjälp för att syntetisera gruppinput eller genomföra forskning för teamet.

• Tidigare diskussioner och beslut, hjälper till att upprätthålla kontinuitet och sammanhållning inom gruppen.

• Tillgänglighet kan underlätta kommunikationen.

• Det hjälper till att förstå social dynamik som är viktiga aspekter av effektiv gruppkommunikation.

• Det hjälper kreativiteten som när jag kan generera svar baserat på befintliga data och mönster.

• Det hjälper anpassningsförmågan, till exempel samtidigt som jag kan anpassa mig till vissa förändringar i gruppdynamiken.

Framtidsperspektiv

Att reflektera över tidigare erfarenheter och identifiera förbättringsområden är avgörande för att förbättra min effektivitet i framtida grupparbetsscenarier.

3.2 Laddar datauppsättning i R

Först behövdes det för att installera, ladda nödvändiga bibliotek och importera data.

```
> # Load the readxl package
> library(readxl)
>
> # Load Excel file
> car_data <- read_excel("C:/Rumon/DS23EC/6. R/Kunskapskontroll/Final/car_data.xlsx")
> # View dimension and head
> dim(car_data)
[1] 701  12
> head(car_data)
# A tibble: 6 x 12
  Car_Price County  Fuel  Gearbox Mileage Model_Year Car_Type Drivetrain_System
      <dbl> <chr>   <chr> <chr>     <dbl>      <dbl> <chr>    <chr>
1    419000 Stockh… Milj… Automat   14452       2017 SUV      Fyrhjulsdriven
2    329800 Stockh… Dies… Automat   11818       2018 SUV      Fyrhjulsdriven
3    239800 Stockh… Milj… Automat   11188       2016 SUV      Fyrhjulsdriven
```

Fig. 2 Importera data i R

3.3 Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) gjordes för att se kvaliteten på data, hitta saknade värden i data och fixa det.

```
> print(missing_values)
       Car_Price           County             Fuel          Gearbox
               0                0                0                0
         Mileage       Model_Year         Car_Type Drivetrain_System
               0                0                0                0
      Horsepower            Color            Brand            Model
               0                0                0                0
```

Fig. 3 Visar inga kvarvarande saknade värden i data

3.4 Data visualisation

Data visualisation I detta skede genomfördes datavisualisering som är Histogram, Bar plot, Scatter plot och Box plot som hjälper oss att utforska och förstå mönster, trender och insikter i data.

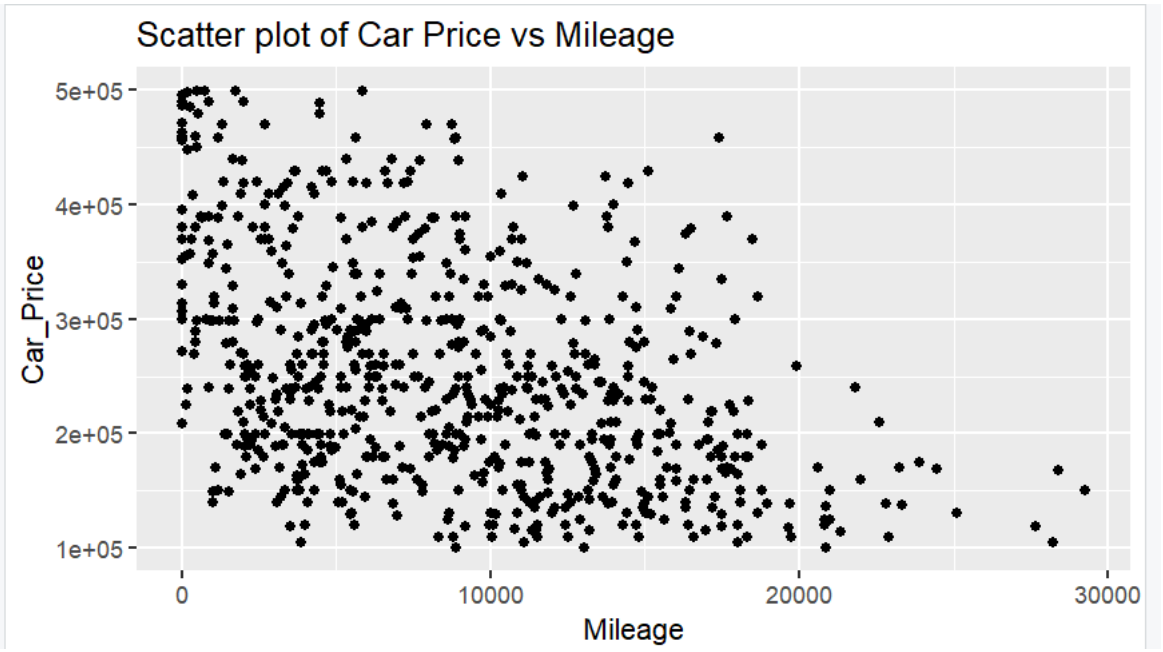

Fig. 4 Scatter plot av bilpris vs körsträcka

3.5 Splitting data into training and test

Före multipel regressionsanalys gjordes det i detta steg att dela upp data i tränings- och testset.

```
# Split data into training and testing sets
set.seed(456)  # For reproducibility
train_index <- sample(seq_len(nrow(car_data)), size = floor(0.7 * nrow(car_data)))
train_data <- car_data[train_index, ]
test_data <- car_data[-train_index, ]
```

Fig. 5 Dela upp data i träning och test

3.6 Multiple Regression Analysis

I detta skede utfördes multipel regressionsanalys och tre olika modeller skapades.

```r
# Build multiple regression models

# Model 1: Using all available predictors
model1 <- lm(Car_Price ~ ., data = train_data)

# Model 2: Using a subset of predictors
model2 <- lm(Car_Price ~ Mileage + Model_Year + Horsepower + Brand, data = train_data)

# Model 3: Using a different subset of predictors
model3 <- lm(Car_Price ~ Fuel + Drivetrain_System + Color + Car_Type, data = train_data)
```

Fig. 6 Regression modeller

3.7 Trained the models and prediction of car price

Efter att ha skapat 3 modeller i regressionsanalys tränades alla modellerna så att de kunde förutsäga värden. Det togs ett specifikt verkligt bilpris från datamängden (419 000) och jämfördes med det förutspådda priset av tre modeller som kan ses nedan

```
print(comparison)
   Model Predicted_Price Original_Price
Model 1         414906.4          419000
Model 2         348389.6          419000
Model 3         355365.6          419000
```

Fig. 7 Förutspådda värden för varje modell

3.8 Confidence Interval (CI)

Konfidensintervall utfördes med 95 % konfidens.

```
print(conf_intervals)
       fit      lwr      upr
414906.4 379623.3 450189.4
```

Fig. 8 Konfidensintervall

3.9 Prediction interval (PI)

Prediktionsintervall utfördes med 95 % konfidens.

```
print(pred_intervals)
       fit      lwr      upr
414906.4 335041.1 494771.6
```

Fig. 9 Prediktionsintervall

4 Resultat och Diskussion

Jämförelse av modeller genom R-squared (R^2)

Model 1 har det högsta R-squared (0,937) av de tre modellerna. Detta indikerar att det förklarar ungefär 93,7 % av variationen i bilpriserna som tyder på en mycket bra anpassning till data.

Model 2, som använder en delmängd av prediktorer (Milage, Model_Year, Horsepower, Brand), har ett lägre R-squared (0,769) jämfört med Model 1 men fångar fortfarande en betydande del av variansen i bilpriser.

Model 3, som använder en annan delmängd av prediktorer (bränsle, Drivetrain_System, Color, Car_Type), har det lägsta R-squared (0,336), vilket indikerar att det förklarar den minsta variationen i bilpriser bland de tre modellerna.

Root Mean Square Error (RMSE)

RMSE-värdena ger en indikation på medelfelet för modellens förutsägelser. Lägre RMSE-värden indikerar bättre prediktiv prestanda.

Model 3 har den lägsta RMSE som tyder på att den har det minsta genomsnittliga prediktionsfelet jämfört med de andra modellerna. Det är dock viktigt att överväga andra utvärderingsmått vid sidan av RMSE.

Adjusted R-squared

Adjusted R-squared tar hänsyn till antalet prediktorer i modellen och straffar för övermontering. Det är vanligtvis lägre än R-squared men ger en mer konservativ uppskattning av modellens förklaringskraft.

Model 1 har fortfarande den högsta adjusted R-squared (0,821) som indikerar att den upprätthåller en bra balans mellan modellens komplexitet och förklaringskraft.

Model 3 har den lägsta adjusted R-squared (0,258) som tyder på att den kan lida av överpassning eller otillräcklig inkludering av prediktorer.

Bayesian Information Criterion (BIC):

BIC är ett kriterium för modellval som straffar modeller med fler parametrar. Lägre BIC-värden indikerar en bättre balans mellan modellpassning och komplexitet.

Model 2 har den lägsta BIC som tyder på att det kan vara den mest lämpliga modellen bland de tre baserat på detta kriterium.

Ytterligare analys, såsom att undersöka betydelsen av individuella prediktorer och genomföra korsvalidering, kan dock hjälpa till att förfina modellvalsprocessen och säkerställa robustheten och generaliserbarheten hos den valda modellen.

5 Slutsatser

Model 1 har gett den bästa övergripande prestandan baserat på de högsta R-squared (R^2) och adjusted R-squared värdena som indikerar en bra balans mellan förklaringskraft och modellkomplexitet. Därför har Model 1 valts ut av de tre modellerna för denna regressionsanalys.

6 Teoretiska frågor

1. Quantile-Quantile (QQ) plot är ett grafiskt verktyg för att hjälpa oss att bedöma om en uppsättning data troligen kom från någon teoretisk fördelning som en normal eller exponentiell. Till exempel, om vi kör en statistisk analys som antar att våra residualer är normalfördelade, kan vi använda en normal QQ-plot för att kontrollera det antagandet. Det är bara en visuell kontroll, inte ett lufttätt prov.

En QQ-plot är en scatterplot skapad genom att plotta två uppsättningar av kvantiler mot varandra. Om båda uppsättningarna av kvantiler kom från samma fördelning borde vi se punkterna som bildar en linje som är ungefär rak.

2. Här är mitt svar till Karin:

Du har rätt Karin! Jag skulle vilja beskriva detta med exempel så att det blir tydligare för dig eftersom du har hört detta men jag har studerat detta :)

Machine Learning (ML), det primära fokus ligger ofta på att bygga modeller som kan göra prediktioner baserat på indata. Anta att vi har en datauppsättning av bostadspriser med funktioner som kvadratmeter, antal sovrum och plats. Med hjälp av dessa data kan en maskininlärningsmodell tränas för att förutsäga priset på ett hus baserat på dessa funktioner. Fokus här ligger på att exakt förutsäga priset på hus för nya datapunkter.

Å andra sidan är statistisk regressionsanalys en metod som används i statistik för att förstå sambandet mellan en beroende variabel och en eller flera oberoende variabler. Förutom att göra prediktioner tillåter regressionsanalys statistisk inferens, vilket innebär att man drar slutsatser om populationen baserat på urvalsdata. Genom att använda samma bostadsdataset kan regressionsanalys användas för att inte bara förutsäga huspriser utan också förstå sambandet mellan varje funktion och priset. Till exempel kan en regressionsanalys avslöja att ytor har en stark positiv korrelation med huspris, medan antalet sovrum har en svagare korrelation.

3. Ett konfidensintervall används i statistisk inferens för att uppskatta det intervall inom vilket en populationsparameter, såsom ett medelvärde eller en regressionskoefficient, sannolikt ligger. Närmare bestämt tillhandahåller den en rad värden inom vilka vi är övertygade om att det sanna parametervärdet ligger med en viss nivå av konfidens.

När vi gör regressionsanalys ger ett konfidensintervall för ett förutsagt värde ett intervall inom vilket vi är säkra på att det sanna medelsvaret för en given uppsättning prediktorvärden ligger. Den kvantifierar den osäkerhet som är förknippad med att uppskatta befolkningens medelsvar.

Ett prediktionsintervall används för att uppskatta det intervall inom vilket en individuell observation från populationen sannolikt kommer att falla. Den tar hänsyn till både den osäkerhet som är förknippad med att uppskatta medelsvaret (som fångas av konfidensintervallet) och den ytterligare variabilitet som är inneboende i individuella observationer.
När vi gör regressionsanalys ger ett prediktionsintervall för ett förutsagt värde ett intervall som vi förväntar oss att en ny observation faller med en viss nivå av konfidens. Den kvantifierar den osäkerhet som är förknippad med att förutsäga ett individuellt svar.

4. I den multipellinjära regressionsmodell betaparametrarna (β) representerar koefficienterna som kvantifierar sambandet mellan varje oberoende variabel (x1,x2, ..., xp) och den beroende variabeln (Y), medan andra variabler hålls konstanta.

Intercept (β0): The intercept term (β0) represents the expected value of the dependent variable (Y) when all independent variables are set to 0.

Varje lutningskoefficienter (β1, β2, ..., βp) representerar förändringen i det förväntade värdet av den beroende variabeln (Y) för en ändring på en enhet i motsvarande oberoende variabel (xi), som innehåller alla andra variabler konstant.

Till exempel, om β1=3, betyder det att för varje ökning av en enhet i x1, ökar det förväntade värdet av Y med 3 enheter, förutsatt att alla andra variabler hålls konstanta.

Omvänt, om β1= -3, betyder det att för varje ökning av en enhet i x1, minskar det förväntade värdet på Y med 3 enheter, förutsatt att alla andra variabler hålls konstanta.

5. Nedan är mitt svar till Hassan:

Vi vet att i statistisk regressionsmodellering är användningen av träning, validering och testuppsättningar en vanlig praxis för modellutvärdering och -val. Däremot är Bayesian Information Criterion (BIC) ett statistiskt mått som används för modellval som innehåller ett straff för modellkomplexitet, och därigenom åtgärdar frågan om overfitting.

Logiken bakom att använda BIC istället för traditionella träning test delnings- eller korsvalideringstekniker är att BIC tillhandahåller ett principiellt sätt att balansera modellpassning och modellkomplexitet. Det straffar modeller med fler parametrar och gynnar enklare modeller som är mindre benägna att överanpassa data.

6. Algoritmen "Best subset selection" syftar till att hitta den bäst passande modellen från alla möjliga kombinationer av prediktorer.

Här är en sammanfattning av stegen:

Nollmodell (Mo): Den börjar med en nollmodell som inte innehåller några prediktorer. Denna modell förutsäger urvalets medelvärde för varje observation. Det fungerar som en baslinje för jämförelse.
För varje k från 1 till p (antal prediktorer):

(a) Den passar alla möjliga kombinationer av modeller som innehåller exakt k prediktorer av de totala p-prediktorerna.

(b) Bland dessa (p välj k) modeller, välj den med minsta restsumman av kvadrater (RSS) eller motsvarande den största R^2. Detta steg syftar till att hitta den bäst passande modellen för varje antal prediktorer.

Välj den bästa modellen

När modeller för alla möjliga antal prediktorer har identifierats (M_o, M_1, ..., M_p), välj en enda bästa modell bland dem.
Detta urval kan baseras på olika kriterier såsom prediktionsfel på en valideringsuppsättning, Cp (Mallows' Cp), Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), justerad R^2 eller korsvalidering.

Sammanfattningsvis söker algoritmen uttömmande igenom alla möjliga delmängder av prediktorer och utvärderar varje delmängds modellanpassningskvalitet baserat på ett valt kriterium. Den väljer sedan den bästa modellen bland dessa delmängder, med hänsyn till en valideringsuppsättning eller andra modellvalskriterier för att undvika överanpassning.

7. "All models are wrong" vilket säger oss att ingen modell representerar verkligheten perfekt. AI-modeller, precis som statistiska modeller, är förenklade representationer av en mycket mer komplex verklighet. De bygger på antaganden, tränas på delmängder av data och gör förutsägelser baserat på mönster de har lärt sig. Som sådan kan ingen modell perfekt representera varje aspekt av den verkliga världen.

"Some are useful" - vi har redan vetat att modeller inte är helt korrekta, men trots sina begränsningar kan modeller ge oss värdefulla insikter, göra prediktioner, informera beslutsfattande och hjälpa oss att förstå mönster och samband i data. De är verktyg som låter oss interagera med, manipulera och förstå komplexiteten i världen.

7 Extern dataanalys

I detta skede av mitt projekt har jag analyserat extern data. Data analyserades från Statistiska Centralbyrån (SCB) som är den svenska statens myndighet som lyder under Finansdepartementet och ansvarar för att ta fram officiell statistik för beslutsfattande, debatt och forskning.

Eftersom min uppgift i min grupp var att samla in bildata från www.blocket.se för Skåne län, här har jag även analyserat skånsk bildata från SCB (https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__TK__TK1001__TK1001A/PersBilarA/). Jag har analyserat från SCB:s hemsida "Personbilar i trafik efter län och kommun samt ägande, År 2002 - 2023."

Nedan följer data som jag har analyserat:

År	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Skåne län											
Antal bilar	517569	522636	529049	537687	548832	559398	563246	569202	574107	581199	586880

År	2013	2014	2015	2016	2017	2018	2019	2020	2021	2022	2023
Skåne län											
Antal bilar	594937	604755	614845	629359	639700	645028	651510	660896	666617	663102	662586

Tabell 1. Totalt personbilar i trafik i Skåne län från 2022 till 2023

Min analys av ovanstående data

Om vi ser stapeldiagrammet och linjediagrammet för datan blir det tydligare för oss hur datamönstret är. Visualisering hjälper oss att förstå data bra.

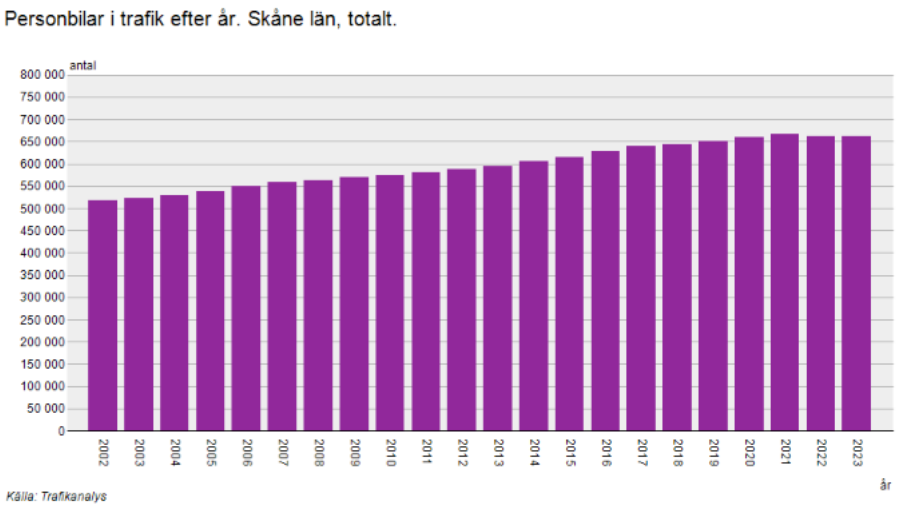

Fig. 10 Stapeldiagram av personbilar i trafik i Skåne län från 2022 till 2023

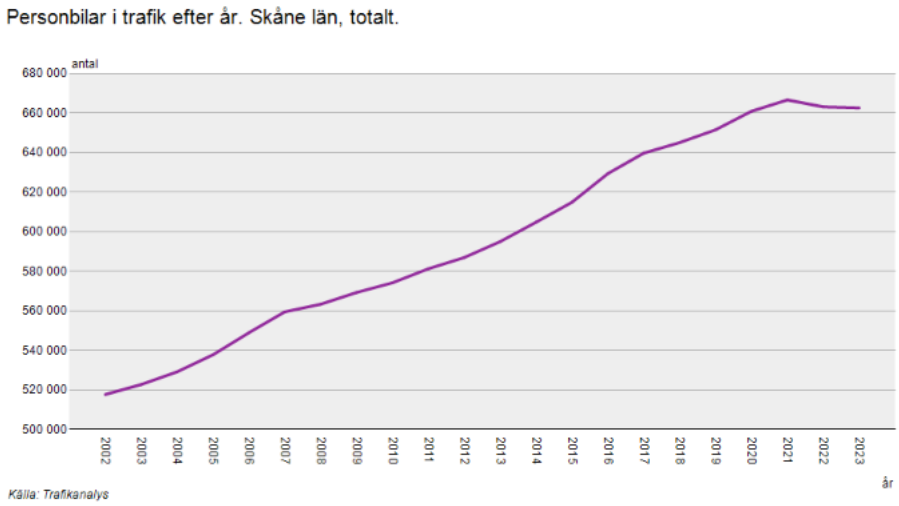

Fig. 11 Linjediagram av personbilar i trafik i Skåne län från 2022 till 2023

Nedan är mina resultat från ovanstående data:

- Mean number of cars: 596506.4
- Median number of cars: 590908.5
- Mode number of cars: 517569
- Range of cars: 149048
- Standard Deviation of cars: 50058.58
- Variance of cars: 2505861344

Efter att ha gjort linjär regressionsanalys:

- Residual Standard Error: 5653
- R-squared (R^2): 0,9879
- Adjusted R-squared: 0,9872
- F-statistics: 1627
- p-value: < 2,2e-16
- Predicted number of cars for the year 2024: 684619
- Confidence Interval (95%): [672827 , 696411]

Ovanstående statistik indikerar att den linjära regressionsmodellen ger en bra passform till data, med antalet bilar som visar en signifikant trend över åren. Modellen förklarar en stor del av variansen i antal bilar och sambandet mellan årtal och antal bilar är mycket signifikant.

Sammanfattning:

■ Det förväntade antalet bilar för 2024 uppskattas till 684619 genom linjär regressionsmodell som överstiger både medelvärdet för bilar och det maximala antalet bilar som observerats under ett enskilt år i datasetet.

■ Konfidensintervallet på 95 % för det prognostiserade antalet bilar år 2024 beräknas vara mellan 672827 och 696411 som indikerar en hög nivå av tilltro till det prognostiserade värdet.

Slutsats:

➤ Analysen tyder på en markant uppåtgående trend av det totala antalet personbilar i trafiken inom Skåne län. Det prognostiserade antalet bilar för 2024 är högre än något tidigare år som understryker den kontinuerliga tillväxten i bilägande och användning inom regionen.

➤ Denna prognos har betydande konsekvenser för olika intressenter, inklusive stadsplanerare, transportmyndigheter och miljöpolitiska beslutsfattare. Den beräknade ökningen av antalet bilar understryker behovet av hållbar transportplanering, utveckling av infrastruktur och åtgärder för att hantera trafikstockningar och miljöpåverkan.

➤ För att effektivt hantera den växande efterfrågan på privata fordon är det absolut nödvändigt för beslutsfattare att prioritera investeringar i kollektivtrafik, främja alternativa transportsätt och implementera policyer som syftar till att minska bilberoendet och främja hållbara mobilitetslösningar.

8 API till extern data

API gjordes med den givna url:n: https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/TK1001/TK1001A/PersBilarA i detta skede av projektet.

Det användes Python-programmeringsspråket för att etablera detta API. Datan som vi kan se på hemsidan (https://www.statistikdatabasen.scb.se/pxweb/sv/ssd/START__TK__TK1001__TK1001A/PersBilarA/) och i kapitel 7 Extern dataanalys, nu kan den ses i Jupyter Notebook genom att köra koden. Vi kan se hur det ser ut från nedanstående bilder:

```
# Define the URL
api_url = "https://api.scb.se/OV0104/v1/doris/sv/ssd/START/TK/TK1001/TK1001A/PersBilarA"

# Call the function with the provided URL
result = fetch_data_from_api(api_url)
print(result)
```

```
{"title":"Personbilar i trafik efter region, ägarkategori, tabellinnehåll och år","variabl
on","values":["00","01","03","04","05","06","07","08","09","10","12","13","14","15","16","
3","24","25","0114","0115","0117","0120","0123","0125","0126","0127","0128","0136","0138",
3","0180","0181","0182","0183","0184","0186","0187","0188","0191","0192","0305","0319","03
382","0428","0461","0480","0481","0482","0483","0484","0486","0488","0509","0512","0513","
```

Fig. 12 Extern data genom API

```
# Print the DataFrame
print(df)
```

	County	Year	Number of Cars
0	Skåne	2002	517569
1	Skåne	2003	517569
2	Skåne	2004	517569
3	Skåne	2005	517569
4	Skåne	2006	517569
5	Skåne	2007	517569
6	Skåne	2008	517569
7	Skåne	2009	517569
8	Skåne	2010	517569
9	Skåne	2011	517569
10	Skåne	2012	517569

Fig. 13 Antal bilar i Skåne län efter år

9 Självutvärdering

1. Utmaningar du haft under arbetet samt hur du hanterat dem.

Det tog lång tid för mig att skapa tre olika regressionsmodeller för car_data som har samlats in från www.blocket.se. Det visade fel flera gånger när jag körde modellerna. Sedan gick jag igenom lektionsmaterial, Antonios YouTube-video, diskuterade med mina klasskollegor och till slut var problemet löst.

2. Vilket betyg du anser att du skall ha och varför.

Precis som andra kurser jobbade jag hårt även på denna kurs. Jag lärde mig massor av saker från lektioner, diskussioner och grupparbeten. Det här kunskapskontroll inklusive rapportskrivning, kod, extern dataanalys, teoretiska frågor gick väldigt bra även jag har gjort linjär regressionsanalys i extern data av personbilar i trafik för Skåne län insamlad från Statistiska centralbyrån (SCB) för att få en bättre förståelse av kursen även om det inte var ett krav att göra det för kunskapskontrollen. Dessutom har jag gjort VG-delen som skapar API genom att använda Python-programmeringsspråket.

Jag förtjänar VG eftersom ovan nämnda faktorer :)

3. Något du vill lyfta fram till Antonio?

Som alltid är jag mycket nöjd och lärde mig så många nya saker i R programmering för dataanalys som definitivt kommer att hjälpa mig i min professionella karriär. Jag är mycket tacksam mot Antonio Prgomet och ger honom massor av respekt och stort tack!

Källförteckning

1. https://www.sciencedirect.com/topics/social-sciences/multiple-regression-analysis#:~:text=Multiple%20regression%20is%20a%20statistical,of%20the%20single%20dependent%20value.

2. http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/

3. https://www.kevinwangstats.com/post/2021-05-30-confidence-and-prediction-intervals/

4. https://www.ibm.com/topics/api