Maskininlärning och regressionsanalys av pris, försäkring och skatt för begagnade bilar

En statistisk modell baserad på svenska andrahandsbilar

Kunskapskontroll i kursen Programmering i R, Maria Lagerholm
Data Science-programmet, EC Utbildning
202504

Abstract

This project aims to predict the price of used cars in Sweden by combining statistical inference and machine learning. Data was collected automatically from Blocket (online ads), Transportstyrelsen (vehicle attributes), SCB (market trends), and If (insurance costs). By integrating technical specifications, policy-driven factors like tax changes, and economic context, we built a predictive model using Random Forest regression. The model identifies key price drivers, such as horsepower, mileage, and fuel type, and highlights discrepancies between market price and predicted value. Our results provide insights into which assumptions about car pricing hold true and offer a data-driven tool for value estimation.

Innehållsförteckning

Inledning

Det här projektet syftar till att genom statistisk inferens och prediktiv modellering förstå vilka faktorer som påverkar priset på begagnade bilar i Sverige. Målet är att identifiera mönster i marknaden och skapa en modell som kan förklara och förutsäga vad som driver värdet på en bil. Vi vill ta reda på om vanliga uppfattningar stämmer: är elbilar verkligen billigare att försäkra? Spelar motorstyrka någon roll för skatten? Och varför kostar vissa bilar mycket mer än andra?

För att undersöka detta kombineras flera datakällor. Från **Blocket** (blocket.se) hämtas information om annonserade priser och bilars grundegenskaper. Med hjälp av bildigenkänning identifieras **registreringsnummer** (platerecognizer.com), vilket gör det möjligt att koppla varje bil till data från **Transportstyrelsen** (https://fordon-fu-regnr.transportstyrelsen.se/UppgifterAnnatFordon/Fordonsuppgifter) – som bränsleförbrukning, utsläpp och skatt. Dessutom används försäkringsbolaget **If:s** formulär (https://www.if.se/privat/partner/nordea/forsakringar/fordon/bil) för att uppskatta försäkringskostnaden. All data samlas automatiskt in med Python, för att minska risken för mänskliga fel och skapa en repeterbar process.

Utöver detta används statistik från **Statistiska centralbyrån (SCB)** för att förstå hur marknaden utvecklats över tid. Vi tittar särskilt på trender i nyregistreringar av bilar under de senaste tio åren, vilket hjälper oss tolka vad som påverkar tillgång och efterfrågan – till exempel skatteändringar, pandemieffekter eller skiften i konsumentbeteende. Denna kontext blir viktig när vi tolkar modellen och avgör vad som driver prissättningen på begagnade fordon i Sverige.

1 Bakgrund

Att förstå vad som styr priset på begagnade bilar i Sverige handlar inte bara om hästkrafter, årsmodell och körsträcka. Det krävs ett bredare perspektiv – ett där både teknik, policy och konsumentbeteende spelar in. Målet i detta projekt är att bygga en modell som kan förklara hur olika faktorer påverkar värdet på andrahandsmarknaden, och samtidigt testa om vanliga uppfattningar faktiskt stämmer.

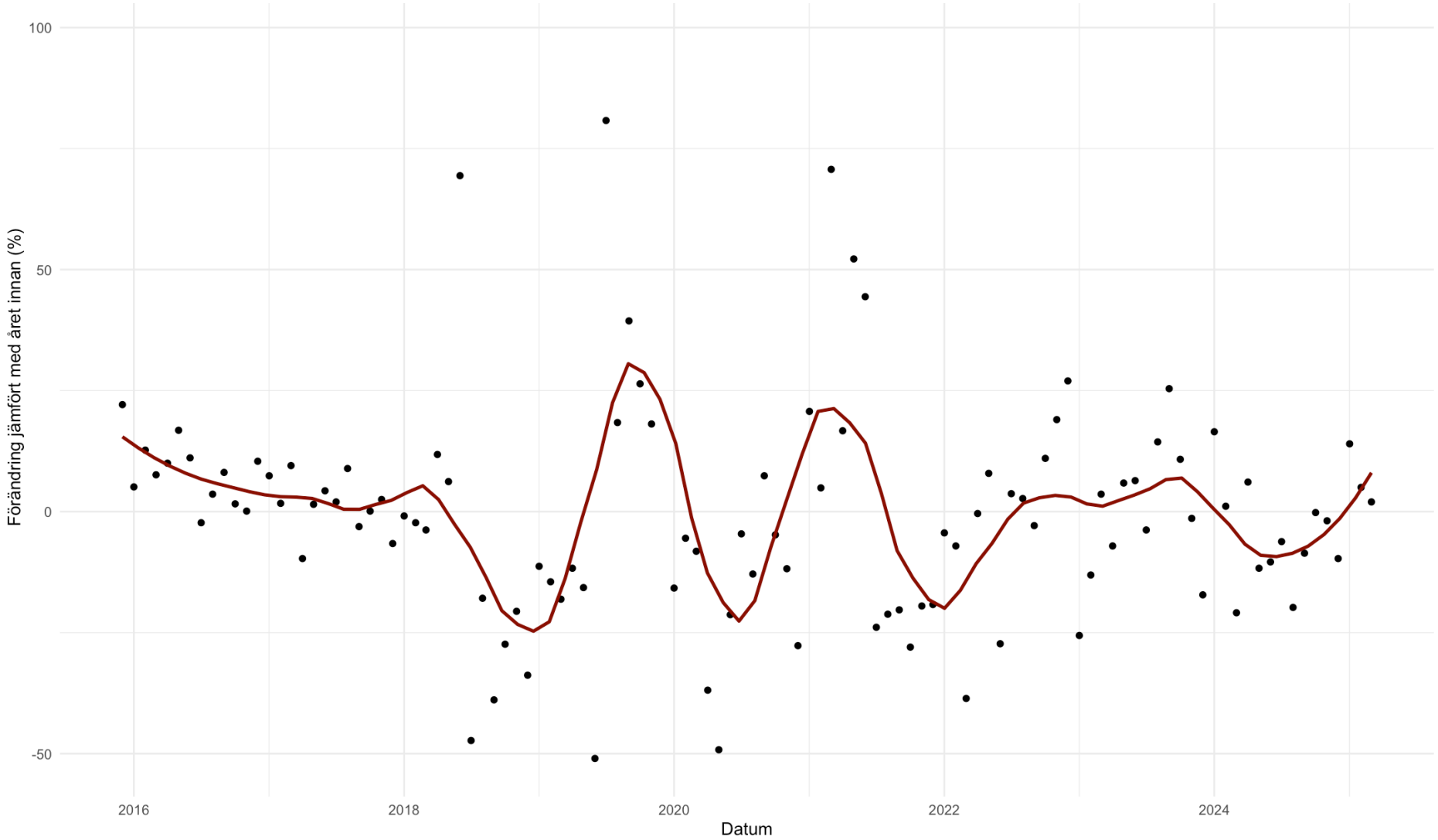
För att göra det har vi kombinerat data från flera centrala källor. Blocket, Sveriges största marknadsplats för begagnade bilar, ger oss pris, modell, bränsletyp och körsträcka – data som samlats in via web scraping. Genom att identifiera registreringsnummer från bilbilder har varje annons dessutom kunnat kopplas till detaljerad fordonsinformation från Transportstyrelsen, såsom utsläpp, skatt och bränsleförbrukning.

Men för att förstå vad som driver marknaden krävs också ett större sammanhang. Därför används statistik från Statistiska centralbyrån (SCB) som speglar samhällsförändringar, till exempel hur många bilar som nyregistreras varje månad. **Figur 1** visar tydliga toppar i nyregistreringar vid årsskiften, särskilt inför förändringar i Bonus–Malus-systemet. När skatteregler ändrades – som i december 2019 och januari 2020 – ökade antalet registrerade bilar markant. Detta beteende återkom flera gånger under 2019–2022 och visar hur starkt politiska beslut kan påverka marknaden.

Sådana förändringar påverkar även begagnatmarknaden. Ett ökat utbud av vissa biltyper kan pressa priser, medan andra blir mer eftertraktade. Därför är det viktigt att förstå hur både bilens egenskaper och omvärldsfaktorer samverkar.

För att komplettera bilden har vi också samlat in försäkringskostnader från If Skadeförsäkring. Dessa påverkas av faktorer som inte alltid syns i en annons – till exempel bilens säkerhetsnivå, skadehistorik eller riskklassificering – men utgör en viktig del av bilens totala ägandekostnad.

Genom att kombinera dessa olika datakällor kan vi analysera priset på begagnade bilar på ett mer nyanserat sätt. Vi kan testa vad som faktiskt har effekt – är det skatten, körsträckan, motorn eller kanske något helt annat?



Figur 1. Trend för förändring i nyregistrerade bilar i Sverige (SCB)

2 Metod

2.1 Statistikmyndigheten SCB

För att förstå strukturella skiften på bilmarknaden hämtades månadsvis statistik över nyregistrerade personbilar från SCB:s öppna API med hjälp av R-paketet pxweb. Vi valde procentuell förändring jämfört med samma månad året innan, från december 2015 till mars 2025. Datan konverterades till data.frame och visualiserades som en tidsserie för att identifiera tydliga toppar kopplade till skatteförändringar. Denna kontext användes för att tolka effekter på andrahandspriser (Mans Magnusson and Markus Kainu and Janne Huovari and Leo Lahti, 2025).

2.2 Datainsamling och dataförberedelse

Datainsamlingen genomfördes med hjälp av en web scraper byggd i Python med Selenium. Verktyget automatiserade insamlingen av bilannonser från Blocket.se, med ett filter för bilar från 2015 och framåt som såldes av privatpersoner. Programmet navigerade genom alla resultatsidor, samlade länkar till annonser, och sparade sedan varje bils pris, titel, specifikationer samt bilder i individuella mappar. Redan hämtade annonser identifierades och hoppades över, vilket möjliggjorde effektiv uppdatering av datamängden. Totalt samlades över 6000 länkar in, varav drygt 800 bilannonser skrapades och sparades strukturerat för vidare analys.

I nästa steg användes en Python-baserad lösning för att automatiskt identifiera registreringsskyltar från bilannonsernas bilder. Genom att integrera med Plate Recognizer API skickades varje bild i respektive bils mapp till tjänsten för skyltigenkänning. Om en giltig skylt identifierades, sparades den i en regnum.txt-fil. Därefter rensades alla bilmappar som saknade denna fil, eller innehöll ogiltiga skyltar (icke-alfa-numeriska eller med fel längd), för att säkerställa att endast annonser med pålitliga registreringsnummer behölls i datasettet. Detta steg var nödvändigt för att möjliggöra vidare insamling av fordonsdata från Transportstyrelsen.

Insamlingen av teknisk fordonsdata från Transportstyrelsens webbtjänst automatiserades genom registreringsnummer. Med hjälp av ett skript i Python och verktyget undetected-chromedriver navigerade vi till varje bils sida, expanderade relevanta sektioner (såsom miljöinformation, teknisk data och skatt), och extraherade nyckeluppgifter som bränsleförbrukning, CO_2-utsläpp, årlig fordonsskatt, bilmärke samt handelsbeteckning. Dessa uppgifter sparades sedan lokalt i respektive fordonsmapp som textfiler. Skriptet var designat för att undvika onödiga anrop genom att endast hämta data för bilar som saknade dessa filer eller där bränsledata tidigare angivits som "0". Tack vare detta steg kunde vi utöka vår dataset med validerad fordonsinformation för vidare analys.

Insamlingen av försäkringskostnader för varje bil automatiserades genom att hämta prisuppgifter från försäkringsbolaget If:s webbsida. För varje bilmapp med ett registreringsnummer (i regnum.txt) navigerade skriptet till If:s formulärsida, fyllde i bilens registreringsnummer och ett personnummer, och klickade på knappen *"Se ditt pris"*. Om ett månadskostnadspris hittades sparades det som text i filen forsakring.txt i respektive bils mapp. Om inget pris kunde hämtas tolkades det som misslyckad identifiering, och mappen raderades. Genom denna metod kunde vi utöka datamängden med ungefärliga försäkringskostnader per månad, vilket är en viktig komponent vid beräkning av bilens totala ägandekostnad.

Sedan skapades en strukturerad dataset i Excel-format från insamlade textfiler i varje fordonsmapp. För varje bil extraherades relevanta variabler såsom registreringsnummer, bränsleförbrukning, skatt, hästkrafter och försäkringskostnad från .txt-filer. Vissa värden omvandlades (exempelvis försäkring per månad till per år och miltal till kilometer). Datasetet sparades som en Excel-fil där varje rad motsvarar en annons, vilket möjliggör vidare analys.

2.3 Statistisk inferens

I inferens delen användes ett flertal R-bibliotek för att möjliggöra statistisk inferens och analys. Bland dessa fanns tidyverse för datahantering och visualisering, readxl för inläsning av Excel-filer, fastDummies för skapande av dummyvariabler, samt corrplot, GGally och ggcorrplot för korrelationsanalyser. För att dela upp data i tränings- och testmängder användes rsample, medan car tillhandahöll verktyg för diagnostik, exempelvis VIF för att upptäcka multikollinearitet. yardstick användes för att beräkna utvärderingsmått såsom RMSE, och ranger, xgboost samt glmnet användes för olika typer av prediktiva modeller. Variabelns betydelse visualiserades med hjälp av vip (Greenwell & Boehmke, 2020).

Vid den statistiska inferensen testades flera grundläggande antaganden för linjär regression. Dessa inkluderade att residualerna skulle vara normalfördelade, vilket kontrollerades med Q-Q-plots och histogram; att variansen hos residualerna skulle vara konstant (homoskedasticitet), vilket bedömdes genom residualplots; samt att sambandet mellan oberoende variabler och den beroende variabeln var linjärt (Almeida et al., 2019). För att identifiera multikollinearitet användes VIF, där höga värden indikerar att variabler är starkt korrelerade. Slutligen kontrollerades frånvaro av inflytelserika outliers med hjälp av Cook's distance och andra residualdiagnoser (Kim, 2017). Dessa steg säkerställde att regressionsmodellerna uppfyllde de grundläggande statistiska förutsättningarna och därmed kunde användas för tillförlitliga tolkningar.

Effekterna för varje regressionskoefficient har prövats med tvåsidiga t-test där nollhypotesen $\beta = 0$ ställs mot alternativet $\beta \neq 0$, vilket ger de p-värden och konfidensintervall (Ismay et al., 2025).

I det statistiska inferenssteget analyserade vi samband mellan fordonsvariabler och tre viktiga målvariabler: **helförsäkringskostnad**, **pris** och **fordonsskatt**. För att möjliggöra linjära regressionsmodeller log-transformerades alla snedfördelade variabler. Därefter byggdes separata multipla regressionsmodeller för varje målvariabel, där prediktorer som modellår, hästkrafter, mätarställning och bränsletyp ingick. Resultaten från modellerna tolkades via regressionskoefficienter, konfidensintervall och p-värden för att bedöma vilka faktorer som har signifikant effekt. Modellernas antaganden granskades med hjälp av residualdiagnostik och multikollinearitet kontrollerades med VIF-värden. Detta möjliggjorde tolkning av prediktiva samband och statistisk signifikans i relation till olika bilkostnader.

2.4 Prediktiv modellering

I denna delen användes prediktiv modellering i syfte att skapa en tillförlitlig modell för att förutsäga priset på begagnade bilar baserat på tekniska och ekonomiska variabler. Målet med modellen är dels att kunna ge en generell uppskattning av marknadspriset för ett fordon, dels att identifiera individuella bilar som verkar vara över- eller underprisade i förhållande till modellens förväntan.

Metoden utgick från en datamängd innehållande bland annat pris, försäkringskostnad, fordonsskatt, hästkrafter, mätarställning och bränsleförbrukning. Flera av de numeriska variablerna transformerades med hjälp av naturliga logaritmer för att minska skevhet och

stabilisera variansen. För att minska påverkan från extrema värden togs outliers bort utifrån det 1:a och 99:e percentilet för varje log-transformerad variabel. Därefter delades datan upp i en tränings- och testmängd (80/20) stratifierat på logaritmerat pris.

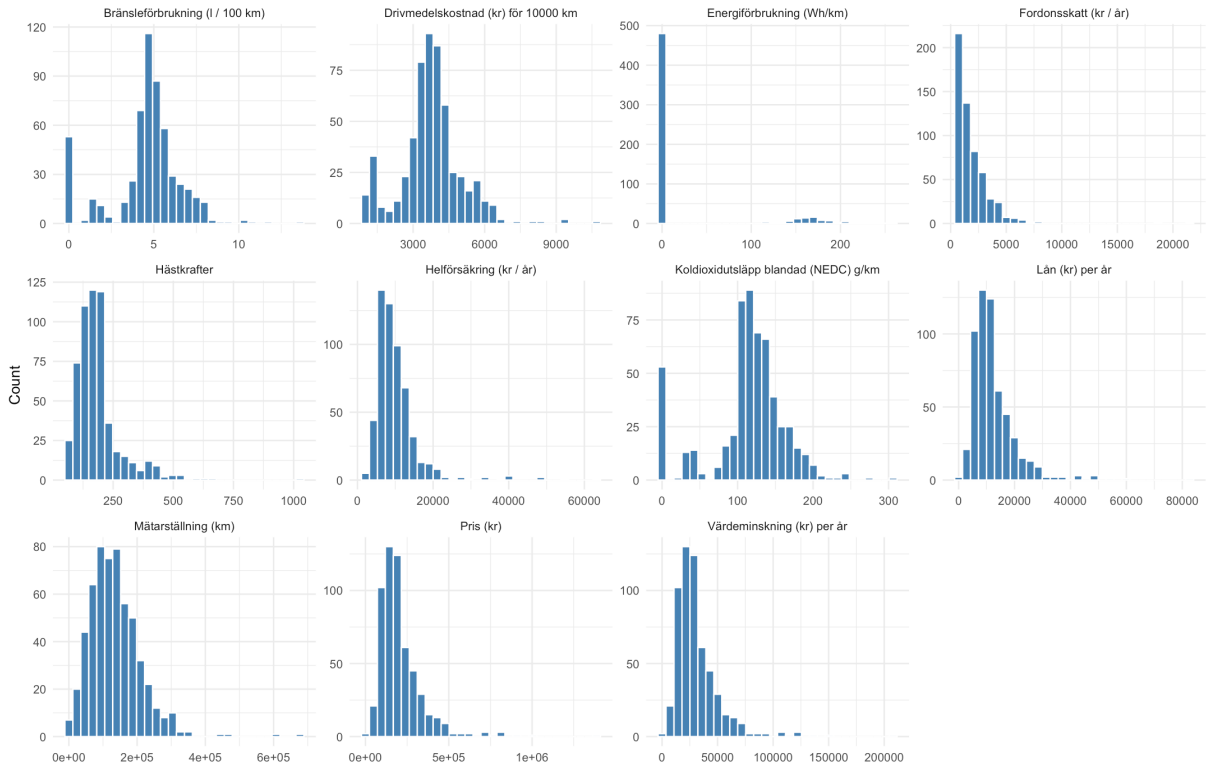
Modellen som användes var en Random Forest-regression, vald för dess robusthet och förmåga att hantera icke-linjära samband och interaktioner utan krav på normalfördelning. Modellens hyperparametrar optimerades med hjälp av grid search i kombination med 5-faldig korsvalidering.

Den slutliga modellen tränades på hela träningsmängden med de bästa inställningarna, och användes därefter för att förutsäga priser på testmängden. Prediktionernas noggrannhet utvärderades med hjälp av root mean squared error (RMSE), både i log-skala och konverterat tillbaka till svenska kronor. Residualanalys genomfördes också för att identifiera bilar vars annonserade pris avvek markant från modellens bedömning.

3 Resultat och Diskution

3.1 Statistisk inferens för försäkringskostnad

Histogrammen i Figur 2 visade tydlig högerskevhet i pris, fordonsskatt, värdeminskning, lån, mätarställning och hästkrafter, vilket motiverade log-transformation för att linjärisera sambanden och dämpa extremvärden. Efter dessa transformationer förklarar den linjära modellen drygt 60 % av variationen i den log-transformerade försäkringspremien ($R^2 = 0.6089$; justerat $R^2 = 0.5982$) och är statistiskt säkerställd med $F(15, 551) = 57.19$, $p < 2{,}2 \cdot 10^{-16}$.



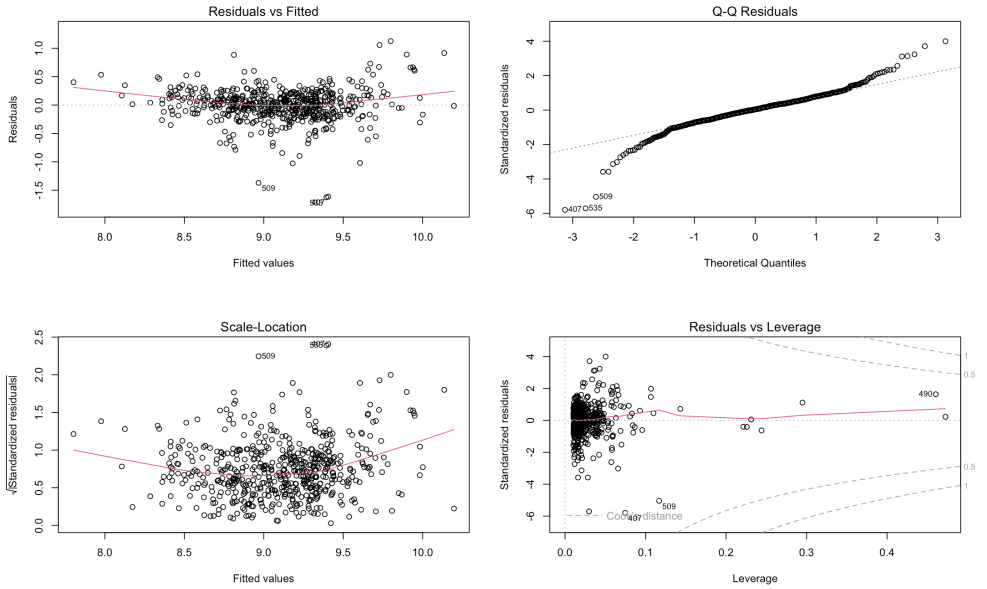
Figur 2. Histogram över centrala fordonsvariabler

Tabell 1. Tolkning av regressionskoefficienter med målvariabel försäkringspris (log-linjär modell), 95 % konfidensintervall.

Variabel	β-estimat	95 % KI	p-värde	Tolkning*
Hästkrafter (log)	0,67	[0,59; 0,76]	< 0,001	+10 % hk ⇒ ≈ 6,7 % högre premie
Modellår 2016 – 2020†	0,13	[0,05; 0,22]	< 0,01	≈ 14 % dyrare än referensår 2015
Modellår 2022 – 2024†	–0,63	[–0,96; –0,38]	< 0,001	≈ 46 % billigare än 2015
Bränsle Diesel	0,19	[0,13; 0,25]	< 0,001	≈ 21 % dyrare än bensin
Bränsle El	0,14	[0,04; 0,25]	0,007	≈ 15 % dyrare än bensin
Bränsle Miljö/Hybrid	–0,04	[–0,13; 0,05]	0,27	Effekt ej säkerställd
Pris (log)	0,09	[0,02; 0,16]	< 0,01	+10 % pris ⇒ ≈ 0,9 % högre premie
Körsträcka (log)	–0,002	[–0,03; 0,03]	0,62	Ingen tydlig effekt

***** Responsvariabeln är log-transformerad; koefficienten kan tolkas som en ungefärlig procentuell förändring i priset. Exempel: β = 0,290 ≈ 29 % → exp(0,290) − 1 ≈ 34 %.
† Värdena för årsgrupperna är medel av de enskilda koefficienterna 2016–2020 respektive 2022–2024

Resultaten sammanfattas i **Tabell 1**, där regressionskoefficienter, 95 procent konfidensintervall och korta tolkningar presenteras. Hästkrafter (log) har den starkaste positiva effekten: en ökning med 10 % i motorstyrka höjer premien med ca 6,7 %. Bilens pris (log) har också ett positivt samband med försäkringskostnaden – en tioprocentig prisökning är förknippad med omkring 0,9 % högre premie. Bilar från modellåren 2016–2020 är genomgående dyrare att försäkra än referensåret 2015, medan nyare bilar från 2022–2024 har signifikant lägre premier, vilket sannolikt återspeglar nya garantier och modern säkerhetsutrustning. Bränsletyp har också betydelse: dieselbilar är cirka 21 % dyrare att försäkra än bensinbilar, medan elbilar är ungefär 16 % dyrare. Däremot visar miljöbränsle/hybridbilar ingen statistiskt säkerställd skillnad. Körsträcka (log) har inte någon tydlig effekt på premien.



Figur 3. Diagnostikplottar för regressionsmodell med försäkringskostnad som målvariabel

Modelldiagnostiken (Figure 5) visar inga systematiska avvikelser i residualerna, god varianshomogenitet och endast ett fåtal outliers med hög leverage. Alla GVIF-värden är < 2, så multikollinearitet är ej ett problem.

3.2 Regressionsanalys av bilpris

Prismodellen förklarar cirka 62 % av variationen i log-priserna ($R^2 = 0{,}617$; justerat $R^2 = 0{,}608$) och är statistiskt säkerställd ($F(14, 552) = 63{,}6$; $p < 2{,}2 \cdot 10^{-16}$).

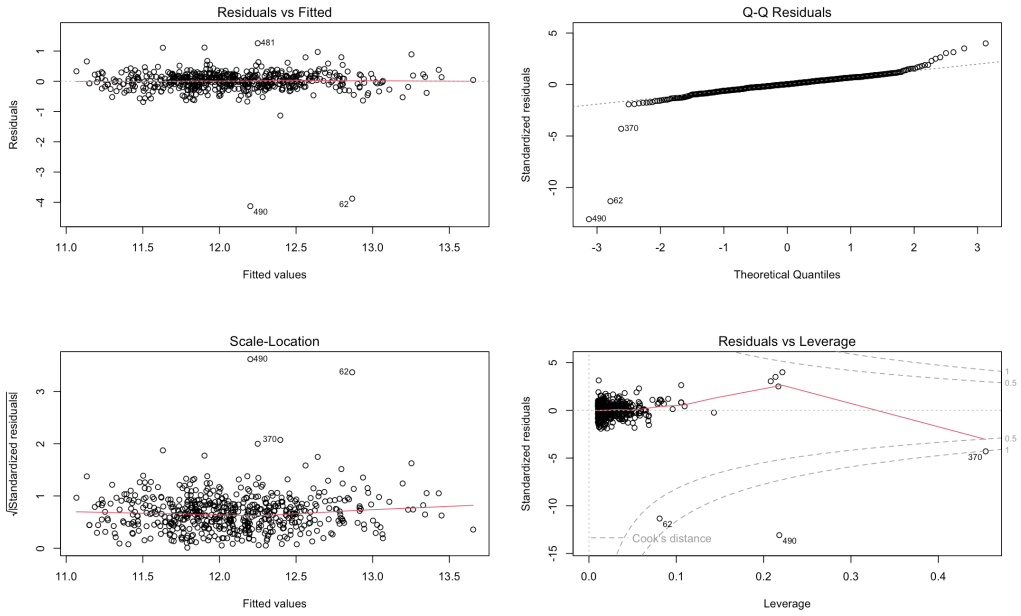
Tabell 2. Tolkning av regressionskoefficienter med målvariabel bilpris (log-linjär modell), 95 % konfidensintervall

Variabel	β estimat	95 % KI	p-värde	Tolkning*
Hästkrafter (log)	0,859	[0,783 ; 0,935]	< 0,001	+1 % fler hk ⇒ ≈ 0,86 % högre pris
Modellår 2016	0,179	[0,081 ; 0,277]	< 0,001	≈ 20 % dyrare än 2015
Modellår 2017	0,295	[0,194 ; 0,396]	< 0,001	≈ 34 % dyrare än 2015
Modellår 2018	0,421	[0,316 ; 0,527]	< 0,001	≈ 52 % dyrare än 2015
Modellår 2019	0,529	[0,412 ; 0,646]	< 0,001	≈ 70 % dyrare än 2015
Modellår 2020	0,581	[0,449 ; 0,712]	< 0,001	≈ 79 % dyrare än 2015
Modellår 2021	0,625	[0,470 ; 0,781]	< 0,001	≈ 87 % dyrare än 2015
Modellår 2022	0,809	[0,627 ; 0,991]	< 0,001	≈ 124 % dyrare än 2015
Modellår 2023	0,553	[0,313 ; 0,793]	< 0,001	≈ 74 % dyrare än 2015
Modellår 2024	–0,007	[–0,357 ; 0,342]	0,97	Ingen effekt (få observationer)
Mätar_log	–0,053	[–0,091 ; –0,014]	0,008	+1 % mer mil ⇒ ≈ 0,05 % lägre pris
Bränsle Diesel	0,035	[–0,039 ; 0,109]	0,35	Ingen säker skillnad mot bensin
Bränsle El	–0,258	[–0,383 ; –0,132]	< 0,001	≈ 23 % billigare än bensin
Bränsle Miljö/Hybrid	–0,095	[–0,211 ; 0,021]	0,11	Ingen säker skillnad mot bensin

* Responsvariabeln är log-transformerad; β≈procentuell prisförändring: $\Delta\% \approx (\exp \beta - 1) \times 100$.

Motorstyrka visar en stark positiv prispremie (**Tabell 2**): en procentuell ökning i hästkrafter ger nästan lika stor procentuell ökning i priset. Nyare årsmodeller (2016–2023) är avsevärt dyrare än 2015, medan 2024 års bilar inte skiljer sig signifikant. Ökad körsträcka är förknippad med en marginell prisminskning. Elbilar är cirka 23 procent billigare än bensinbilar, medan diesel- och hybrideffekterna inte är statistiskt säkra. Diagnostikplottarna (Figur 5) visar lätt tungsvansproblematik och marginell heteroskedasticitet, men inga dominerande observationer eller multikollinearitetsproblem (samtliga GVIF < 1,2), vilket

innebär att modellen är robust och tolkbar för att identifiera de faktorer som driver priset på begagnatmarknaden.



Figur 4. Diagnostikplottar för regressionsmodell med pris som målvariabel

Plottarna (**Figur 4**) visar residualanalys för den linjära modellen med log-transformerat pris. De indikerar i stort sett homogen varians och linjäritet, men vissa avvikelser och inflytelserika observationer (t.ex. ID 490 och 370) bör noteras. Q-Q-plott antyder något avvikande normalfördelning i svansarna. Modellens antaganden är dock i huvudsak rimligt uppfyllda.

3.3 Statistisk inferens för fordonsskatt

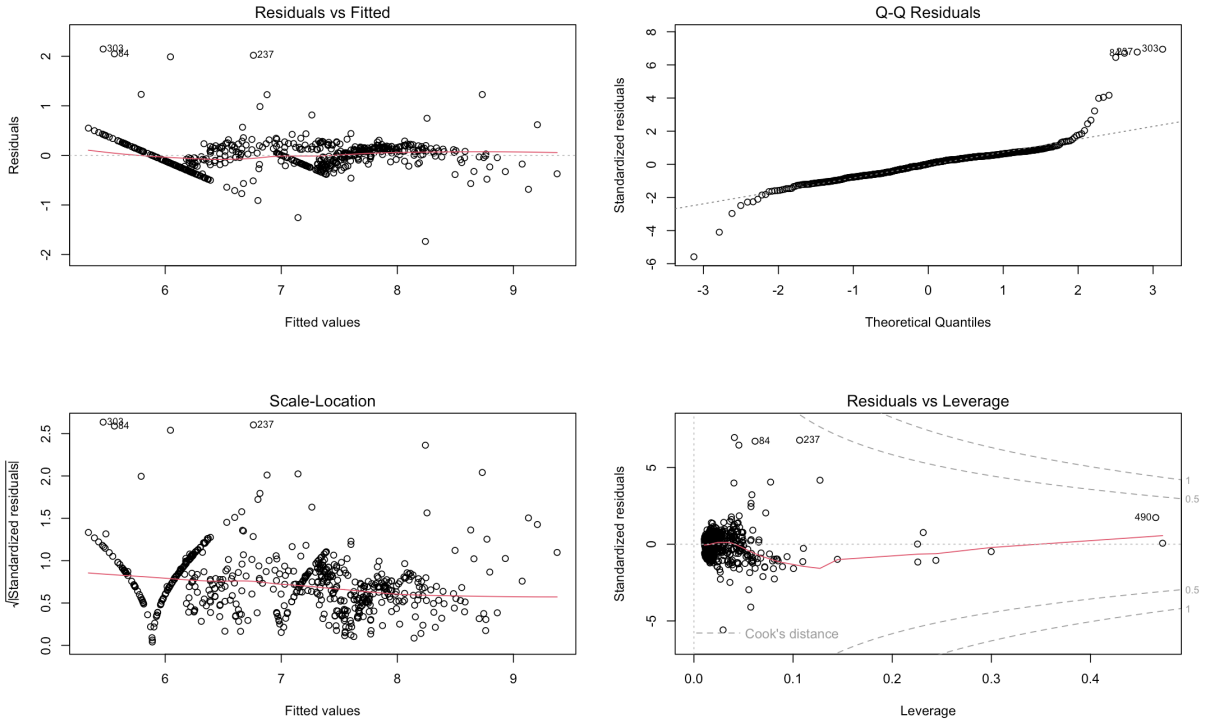
Vi testar varje koefficient med tvåsidiga t-test ($H_0 : \beta = 0$) och 95 % konfidensintervall. Modellen för log-transformerad fordonsskatt förklarar 87,9 % av variationen i skatt ($R^2 = 0{,}8785$; justerat $R^2 = 0{,}8749$) och är mycket statistiskt säkerställd ($F(16, 550) = 248{,}5$; $p < 2{,}2 \cdot 10^{-16}$).

Modellår 2019–2024 är alla signifikanta och innebär att skatten ökar med cirka 13 % för 2019, cirka 47 % för 2020, cirka 52 % för 2021, cirka 83 % för 2022, cirka 71 % för 2023 och cirka 56 % för 2024 jämfört med referensåret 2015 (**Tabell 3**). Bränsletyp har mycket stark effekt: dieselbilar beskattas i genomsnitt cirka 203 % högre och elbilar cirka 141 % högre än bensinbilar. En tioprocentig ökning av bilpriset ger ingen statistiskt säkerställd förändring, medan varje extra gram CO_2 per kilometer höjer skatten med ungefär 1,4 %. Variablerna hästkrafter, körsträcka och miljö-/hybrid visade inte signifikanta samband och bidrar därmed inte statistiskt till skattmodellen.

Tabell 3. Regressionskoefficienter för målvariabeln fordonsskatt (log-linjär modell), 95 % konfidensintervall

Variabel	β estimat	95 % KI	p-värde	Tolkning*
Hästkrafter (log)	0,0844	[–0,0100 ; 0,1789]	0,080	Ingen signifikant effekt
Modellår 2019	0,1371	[0,0268 ; 0,2474]	0,015	≈ 13 % högre skatt än 2015
Modellår 2020	0,3889	[0,2638 ; 0,5140]	< 0,001	≈ 47 % högre skatt än 2015
Modellår 2021	0,4169	[0,2707 ; 0,5631]	< 0,001	≈ 52 % högre skatt än 2015
Modellår 2022	0,6054	[0,4328 ; 0,7780]	< 0,001	≈ 83 % högre skatt än 2015
Modellår 2023	0,5351	[0,3176 ; 0,7525]	< 0,001	≈ 71 % högre skatt än 2015
Modellår 2024	0,4454	[0,1334 ; 0,7573]	0,005	≈ 56 % högre skatt än 2015
Körsträcka (log)	–0,0155	[–0,0502 ; 0,0192]	0,380	Ingen tydlig effekt
Bränsle Diesel	1,0984	[1,0328 ; 1,1639]	< 0,001	≈ 203 % högre skatt än bensin
Bränsle El	0,8812	[0,7020 ; 1,0604]	< 0,001	≈ 141 % högre skatt än bensin
Bränsle Miljö/Hybrid	0,0799	[–0,0443 ; 0,2042]	0,207	Effekt ej säkerställd
Pris (log)	0,0571	[–0,0181 ; 0,1324]	0,136	Ingen säker effekt
Koldioxidutsläpp (g/km)	0,0141	[0,0131 ; 0,0151]	< 0,001	+1 g/km ⇒ ≈ 1,4 % högre skatt

* Responsvariabeln är log-transformerad; β≈procentuell prisförändring: $\Delta\% \approx (\exp \beta - 1) \times 100$.



Figur 5. Diagnostikplottar för regressionsmodell med fordonsskatt som målvariabel

Diagnostikplottarna (**Figur 5**) visar att residualerna är centrerade kring noll utan tydlig kurvatur i Residuals vs Fitted. Här ser vi struktur i residualerna, vilket kan tyda på heteroskedasticitet eller modellfel, t.ex. att viktiga kategorivariabler saknas eller att modellen är för enkel för datans struktur. QQ-plottarna indikerar en något tung högersvans men följer normallinjen för större delen av datan. Scale-Location-diagrammet är i stort sett horisontellt. Det finns indikationer på heteroskedasticitet, särskilt inom vissa delgrupper (t.ex. fordonstyper med olika skattestrukturer). I Residuals vs Leverage syns inga punkter utanför Cook's-distanslinjen 1, vilket tyder på att inga enskilda observationer dominerar modellen. VIF-värdena (GVIF ≤ 2) ligger väl under kritiska nivåer, vilket visar att multikollinearitet inte

är ett problem. Sammanfattningsvis är modellen robust och ger tillförlitliga inferenser om vilka faktorer som driver fordonsskatten.

3.4 Prediktiv modellering av bilpriser med Random Forest

3.4.1 Förberedelse av tränings- och testdata

Först delades den färdigförbehandlade datamängden – bestående av 567 Blocket-bilar – upp i 80 % träningsdata och 20 % testdata. Stratifiering gjordes på den logaritmerade prisvariabeln (pris_log) för att säkerställa att prisfördelningen förblev liknande i båda delmängderna.

Alla kontinuerliga prediktorer – pris, fordonsskatt, helförsäkring, hästkrafter och mätarställning – log-transformerades. Syftet var att minska skevheten i datan och stabilisera variansen, vilket gör det enklare för modellen att identifiera mönster.

Därefter rensades både tränings- och testdata från extrema observationer: rader som låg under första percentilen eller över den 99:e percentilen i någon av de log-transformerade variablerna togs bort. Syftet med detta var att undvika att mycket ovanliga bilar – särskilt exklusiva premiumbilar – skulle få ett oproportionerligt inflytande på trädstrukturen i modellen. Genom denna begränsning kunde modellen fokusera på den typ av bilar som var vanligast förekommande i datan, det vill säga mellanklassbilar i prisklassen under en miljon kronor. En viktig insikt från detta steg var att modellen, tränad på vanliga bilar, inte bör användas för att förutsäga priser på särskilt dyra eller ovanliga fordon. Genom att exkludera dessa outliers förbättrades modellens prestanda markant – RMSE minskade från 0.36 till 0.23 log enheter.

3.4.2 Modellträning och hyperparameteroptimering

Det rensade tränings-settet användes sedan i en femfaldig, stratifierad korsvalidering där en liten hyperparameter-grid för Random Forest utvärderades. Grid-sökningen (**Tabell 4**) omfattade antalet träd, andelen variabler som slumpas i varje delning (mtry), minsta nodstorlek och maximal träddjup. Medel-RMSE (på log-skala) beräknades för varje parameterkombination och den kombination som gav lägst fel valdes för slutmodellen. Tabell 4 sammanfattar den bästa parameteruppsättningen.

Tabell 4. Grid-söknings resultat

trees	mtry	min.node.size	max.depth
50	3	1	12

3.4.3 Prediktionsprestanda och tolkning

Slutmodellen tränades om på hela det rensade träningssettet med ovanstående parametrar och med impurity-baserad variabelvikt aktiverad för att kunna rangordna prediktorerna. På det helt osedda testsettet nådde modellen den prestanda som visas i Tabell 5. Med ett RMSE på log-skalan som ligger mycket nära varandra mellan träning och test (0.230 vs. 0.236), kan resultaten tolkas som att modellen **har god generaliseringsförmåga (Tabell 5)**. För att kunna tolka modellens fel i faktiska kronor (SEK), omvandlades RMSE-värdet från log-skalan till SEK. Eftersom RMSE i log-skala representerar ett relativt fel (ungefär som en procentuell avvikelse), multiplicerades detta värde med det genomsnittliga predikterade priset i SEK på testdatan. Denna metod ger en uppskattning av det genomsnittliga felet i kronor:

$$\text{RMSE}_{SEK} \approx \text{Genomsnittligt predikterat pris} \times \text{RMSE}_{\log}$$

Att använda den naturliga exponentialfunktionen exp(RMSE_log) ger inte ett korrekt resultat, eftersom log-RMSE inte är ett direkt loggat värde som kan "avloggas" med exp() utan representerar ett genomsnitt av kvadrerade fel i log-skala.

$$\mathrm{RMSE}_{\log} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\log(\mathrm{price}_i) - \log(\hat{\mathrm{price}}_i))^2}$$

Eftersom log och kvadrering inte är inverterbara med exponentialfunktionen i detta sammanhang, kan exp(RMSE_log) inte tolkas som ett genomsnittligt absolut fel i kronor. Därför används istället en approximation där log-RMSE multipliceras med det genomsnittliga predikterade priset i SEK för att ge ett mer meningsfullt mått i faktiska pengar.

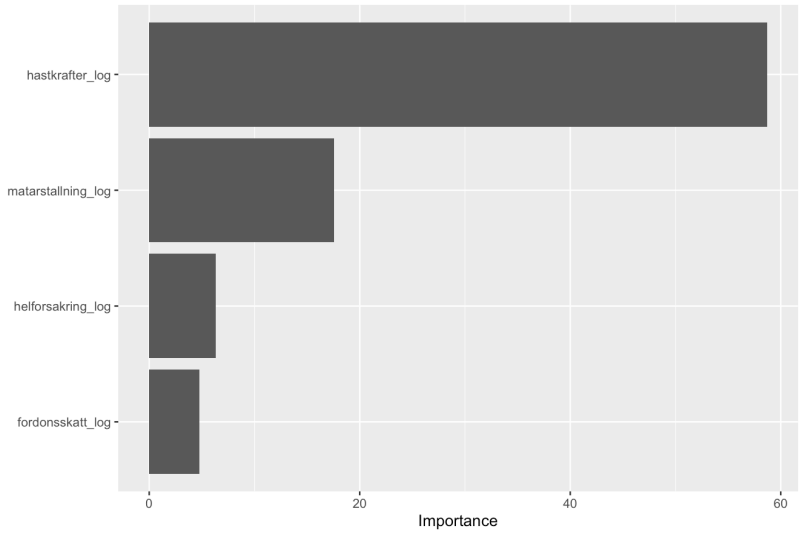
$$\mathrm{RMSE}_{SEK} \approx \bar{y}_{\mathrm{pred}} \times \mathrm{RMSE}_{\log}$$

Till exempel betyder ett RMSE på 0.2313 i log-skala att det genomsnittliga felet är cirka 23 % av priset, vilket motsvarar ungefär 44 297 kr för en bil med genomsnittligt predikterat pris.

Tabell 5. Uppskattning av modellens generaliseringsförmåga

	RMSE (log-skala)	**RMSE (SEK)**
Train (CV K=5)	0.2313	44297
Test	0.2353	56829

Trots att RMSE i log-skala är nästan identisk för träningsdata (0.2313) och testdata (0.2353), är felet i SEK väsentligt större i testuppsättningen (56 829 kr jämfört med 44 297 kr). Detta beror på att testdatan innehåller bilar med ett högre genomsnittligt pris, vilket gör att ett procentuellt fel motsvarar ett större absolut belopp i kronor. Eftersom RMSE i log-skala uttrycker ett relativt fel, påverkas den omräknade SEK-nivån direkt av prisnivån i respektive datamängd.



Figur 6. Variabelvikt enligt Random Forest-modellen för bilpris (log-transf.)

Baserat på variabelimportansdiagrammet från den tränade random forest-modellen framgår det tydligt att **hästkrafter** är den mest avgörande prediktorn för bilens pris, följt av **mätarställning (Figur 7)**. Det är särskilt intressant att mätarställning (antal körda kilometer) tilldelas så hög vikt i random forest-modellen, eftersom detta samband inte fångades upp som

signifikant i den linjära regressionsmodellen under inferensdelen. Denna skillnad tyder på att sambandet mellan mätarställning och pris kan vara **icke-linjärt eller beroende av interaktioner** med andra variabler – något som random forest, till skillnad från linjär regression, kan upptäcka utan att det uttryckligen modelleras. Detta illustrerar hur prediktiva modeller och statistiska modeller ibland identifierar olika mönster i datan, och därmed kompletterar varandra.

3.4.4 Identifiering av avvikande prisnivåer

Tabell 6. Topp 5 Överprisade Bilar

ID	Fordonsbenämning	Pris (kr)	Prognos (kr)	Skillnad (kr)	Skillnad (%)
481	KIA EV9	735 000	420 156	+314 844	+74.9 %
333	BMW X5 XDRIVE45E	560 000	360 145	+199 855	+55.5 %
269	MERCEDES-BENZ V-KLASS E	469 000	329 714	+139 286	+42.2 %
360	SUBARU OUTBACK	305 000	185 964	+119 036	+64.0 %
478	VOLVO XC60	270 000	166 502	+103 498	+62.2 %

Modellen visar tydligt att vissa bilar är kraftigt över- eller underprisade i förhållande till vad den förväntar sig baserat på bilens egenskaper (**Tabell 6**). Bland de mest överprisade finns flera dyrare modeller som KIA EV9 och BMW X5, där priset i annonserna överstiger modellens uppskattade värde med upp till 75 %. Detta tyder på att dessa fordon antingen är felvärderade eller att deras utrustningsnivåer inte fångas av de tillgängliga variablerna.

Tabell 7. Topp 5 Underprisade Bilar

ID	Fordonsbenämning	Pris (kr)	Prognos (kr)	Skillnad (kr)	Skillnad (%)
327	VOLVO V60	329 900	417 784	−87 884	−21.0 %
64	TESLA MODEL 3	270 000	354 375	−84 375	−23.8 %
390	MITSUBISHI OUTLANDER	208 000	284 415	−76 415	−26.9 %
7	BMW X1 XDRIVE20D	164 900	237 757	−72 857	−30.6 %
280	MERCEDES-BENZ AMG CLA 45	285 000	357 659	−72 659	−20.3 %

Å andra sidan identifierades flera bilar som undervärderade, exempelvis Tesla Model 3 och Volvo V60, där det faktiska priset är 20–30 % lägre än modellens uppskattning (**Tabell 7)**. Det innebär att dessa bilar kan vara attraktiva fynd ur ett värdeperspektiv, enligt modellens logik. Dessa resultat visar att modellen inte bara lyckas identifiera prisnivåer, utan också fångar upp avvikelser som kan ha praktisk betydelse för köpare eller säljare.

4 Slutsats

Projektet genomfördes av Maria Lagerholm och Geisol Yissel Urbina. Vi jobbade delade uppgifterna efter kompetens och granskade kod och resultat tillsammans. All datainsamling har varit automatiserad – från att hämta bilannonser, läsa registreringsnummer, till att samla in teknisk information och försäkringskostnader. Slutdatat innehöll över 500 annonser. Modellen vi byggde kan förutsäga bilpriser med ganska bra träffsäkerhet och visar tydliga mönster i begagnatmarknaden. Men modellen gör också ett stort procentuellt fel på ca 23%. Det beror ofta på att viktiga egenskaper som utrustningsnivå, bilens skick eller servicehistorik inte fanns med i datan. Trots det kan modellen hjälpa till att få en snabb överblick och upptäcka bilar som verkar för dyra eller för billiga jämfört med liknande annonser.

5 Teoretiska frågor

1. En Quantile-Quantile (QQ) plot är ett diagram som jämför fördelningen av ett datamaterial med en teoretisk fördelning (oftast normalfördelningen). Om

datapunkterna ligger nära en rät linje i plottet, tyder det på att datan följer den teoretiska fördelningen. QQ-plottar används därför ofta för att kontrollera om residualer i en regressionsmodell är normalfördelade.

2. I maskininlärning fokuserar man främst på att förutsäga resultat så exakt som möjligt. Ex. vilket pris en bil kommer ha, utan att bry sig så mycket om varför. I statistisk regression vill man också förstå vad som påverkar resultatet och hur starkt, alltså göra inferens. Maskininlärning kan säga vad en bil bör kosta, men regression kan förklara hur mycket hästkrafter eller bränsletyp påverkar priset.
3. Konfidensintervall visar osäkerheten kring modellens medelvärdesprediktion – alltså det genomsnittliga förväntade värdet för en viss kombination av variabler. Prediktionsintervall är bredare och visar osäkerheten kring ett enskilt framtida observation, alltså hur mycket en enskild bils pris kan variera.
4. I den Den multipla linjära regressionsmodellen visar varje β-parameter hur mycket Y i genomsnitt förändras när en viss x-variabel ökar med 1 enhet, **om alla andra variabler hålls konstanta**.
5. BIC (Bayesian Information Criterion) hjälper att välja en modell utan att dela upp datan, eftersom den balanserar passform och komplexitet. Men man behöver ändå testdata för att se hur bra modellen funkar på ny data.
6. Best subset selection handlar om att hitta den bästa kombinationen av prediktorer för en linjär modell. Man börjar med nullmodellen (ingen prediktor). Sedan testar man alla modeller med exakt 1, 2, ..., p prediktorer.För varje antal prediktorer k väljer man den modell som ger bäst passform (lägst RSS eller högst R^2). Till sist väljer man den bästa modellen totalt med hjälp av ett kriterium som BIC, AIC, justerat R^2 eller valideringsfel.
7. Modeller förenklar verkligheten och är därför aldrig helt sanna – men de kan ändå vara användbara för att förstå och förutsäga.

6 Självutvärdering

Genom detta projekt har jag i praktiken förstod skillnaden mellan statistisk inferens och prediktiv modellering, och insett vikten av att hantera icke-linjära samband korrekt. Jag har lärt mig vikten av att följa etiska riktlinjer vid datainsamling, men också att kunna hantera situationer där alla modellantaganden inte är uppfyllda och kunna motivera sina val. Arbetet motsvarar nivå för Väl godkänt, då vi hämtade extern data via SCB:s API, automatiserade insamlingen från Blocket och byggde en modell som både förklarar och förutsäger pris med god precision.

7 Källförteckning

Almeida, A., Loy, A., & Hofmann, H. (2019). Ggplot2 compatible quantile-quantile plots in R. *The R Journal*, *10*(2), 248. https://doi.org/10.32614/rj-2018-051

Greenwell, B., & Boehmke, B. (2020). Variable Importance Plots—An Introduction to the vip Package. *The R Journal*, *12*(1), 343. https://doi.org/10.32614/rj-2020-013

Ismay, C., Kim, A. Y., & Valdivia, A. (2025). Simple linear regression. In *Statistical Inference via Data Science* (pp. 115–154). Chapman and Hall/CRC. https://doi.org/10.1201/9781032724546-5

Kim, M. G. (2017). A cautionary note on the use of Cook's distance. *Communications for Statistical Applications and Methods*, *24*(3), 317–324. https://doi.org/10.5351/csam.2017.24.3.317

Mans Magnusson and Markus Kainu and Janne Huovari and Leo Lahti. (2025). *pxweb: R Interface to PXWEB APIs* [Dataset]. https://github.com/rOpenGov/pxweb