Analyzing and Predicting Used Car Prices Using Machine Learning

Ali Khalil

EC Utbildning

R programmering

2024–03

Abstract

This project focused on analyzing and predicting the price of used Volkswagen cars from the years 2000 to 2022. Using a dataset of 1,205 observations, we applied two predictive modeling techniques: multiple linear regression and XGBoost regressor. The performance of both models was evaluated based on key metrics such as Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared. While both models demonstrated strong predictive capabilities, the XGBoost model outperformed the multiple linear regression model, achieving higher scores and better overall performance. Additionally, our analysis revealed that even though the models selected a wide range of features, similar results were achieved by focusing on the most important predictors. These findings suggest that XGBoost is a more effective approach for predicting used car prices, even when working with a limited subset of relevant features.

1 Innehållsförteckning

2 Inledning

Använda bilmarknaden har vuxit betydligt, och många köpare och säljare söker efter pålitliga sätt att uppskatta bilpriser. Att kunna förutsäga priset på en begagnad bil kan vara av stor vikt för både säljare som vill sätta ett konkurrenskraftigt pris och köpare som söker efter rättvärda erbjudanden. Traditionellt har bilpriser beräknats med hjälp av förenklade modeller, men med hjälp av moderna maskininlärningstekniker kan vi skapa mer exakta och dynamiska förutsägelser. Denna studie undersöker och jämför två maskininlärningsmodeller, multipel linjär regression och XGBoost-regressorn, för att förutsäga priset på begagnade Volkswagen-bilar mellan åren 2000 och 2022.

2.1 Syfte

Syftet med denna studie är att analysera och förutsäga priserna på begagnade Volkswagen bilar med hjälp av maskininlärning. Vi syftar till att jämföra prestandan hos två olika modeller, multipel linjär regression och XGBoost, för att undersöka vilken som ger de mest exakta prisprognoserna. Genom att förstå hur dessa modeller fungerar och deras förmåga att hantera olika funktioner och datamängder, strävar vi efter att skapa en modell som kan ge tillförlitliga bilpriser baserat på relevant data.

2.2 Frågeställning

1. Hur väl kan multipel linjär regression och XGBoost förutsäga priser på begagnade Bilar?
2. Vilka variabler har störst inverkan på prissättningen av begagnade bilar?
3. Hur påverkar extrema avvikande observationer (outliers) modellernas prestanda och prediktioner?

3 Extern Data

Den data vi samlade från SCB genom API visar att:
El och hybridfordon blir allt vanligare i Sverige, vilket tyder på en övergång mot grönare transporter. Traditionella bränsletyper (bensin, diesel) dominerar fortfarande men minskar sakta.Det totala fordonsägandet ökar stadigt från 2021 till 2025.

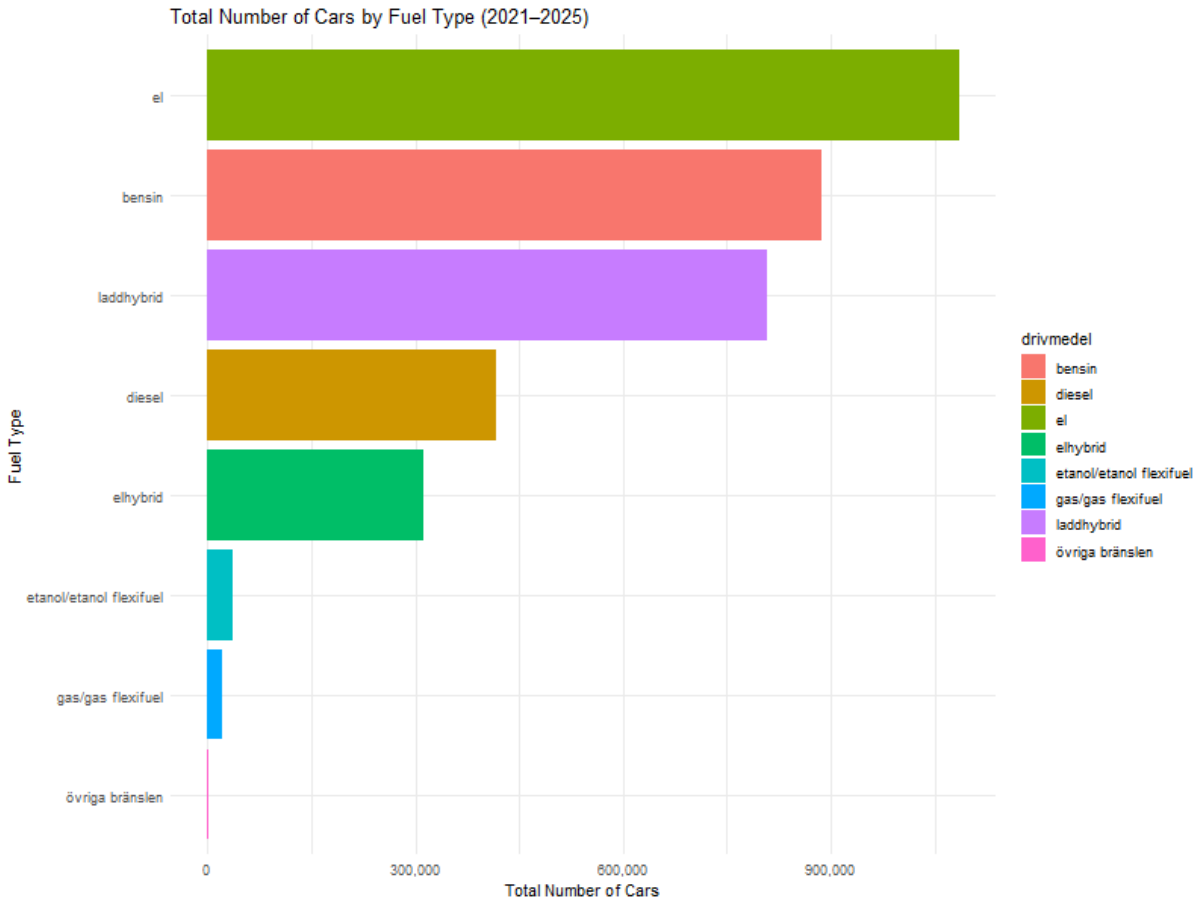


Figure 1:Total number of cars by fuel type2021 to 2025

Även om projektet fokuserar på begagnade Volkswagen-bilar och SCB-data beskriver hela fordonsflottan i Sverige, finns en tydlig koppling:

Den ökande trenden mot hållbarare transporter (t.ex. fler el- och hybridbilar) påverkar begagnatmarknaden och prissättningen av bilar.

Bilar med traditionella bränslen (bensin och diesel), som de flesta Volkswagen-bilar i datasetet troligen är, kan få förändrade värden över tid på grund av den här utvecklingen.

4 Datainsamling

I projektet arbetade vi tillsammans i en grupp bestående av mig, Emil, Andreas, Svetlana, Priya, Camilla, Oskar, Lence, Martin, Gustav, Karl och Andre. Målet var att samla in data om begagnade Volkswagen-bilar från åren 2000 till 2022. Vi bestämde att varje gruppmedlem skulle samla in cirka 100 observationer från Blocket . Totalt lyckades vi samla in 1205 observationer.

För att säkerställa enhetlighet i datan satte vi gemensamma riktlinjer kring vilka variabler som skulle dokumenteras, såsom pris, modell, årsmodell, körsträcka och bränsletyp. Genom tydliga instruktioner kunde vi minska risken för inkonsekvenser och fel vid insamlingen.

Efter insamlingen behövde vi rensa datan från dubbletter och hantera vissa saknade värden innan vi kunde använda den i våra modeller. Vi upptäckte också felstavningar, extra mellanslag och andra små avvikelser i datan som vi var tvungna att korrigera. Detta gjorde att vi blev ännu mer medvetna om vikten av noggrannhet vid manuell datainsamling.

Genom att samla in data manuellt fick vi också en bättre förståelse för hur komplex och rörig verklig data kan vara jämfört med färdigrensade datasets som ofta används i utbildningssammanhang. Erfarenheten betonade vikten av noggrann datarensning och förberedelse innan man går vidare till analys och modellering.

(Blocket, u.d.)

Försäljningspris	Säljare	Bränsle	Växellåda	Miltal	Modellår	Biltyp
179 900	Företag	Bensin	Automat	13800	2017	SUV
165 000	Privat	Miljöbränsle/Hybrid	Automat	18223	2017	Kombi
199 000	Företag	Diesel	Automat	7500	2016	Kombi
229 900	Företag	Diesel	Manuell	15684	2016	Familjebuss
189 900	Företag	Bensin	Automat	14909	2022	Kombi
55 500	Privat	Diesel	Automat	21391	2011	Kombi

Figur 2:exempel åv insamlade data(blocket)

5 Teori

Vid analys och prediktion av priser på begagnade bilar används statistiska och maskininlärningsbaserade modeller för att identifiera sambandet mellan bilens egenskaper och dess pris. I detta projekt har två modeller tillämpats: en multipel linjär regressionsmodell och en XGBoost-regressionsmodell.

5.1 Explorativ Dataanalys (EDA)

Explorativ dataanalys (EDA) är en viktig första fas i dataanalys där syftet är att förstå datans struktur, identifiera mönster, samband, och eventuella avvikelser. Genom EDA kan man få insikter om vilka variabler som kan vara relevanta för modellen och upptäcka problem som saknade värden, felaktiga inmatningar eller outliers. I detta projekt användes EDA för att undersöka sambanden mellan variabler och hur dessa kan påverka bilens pris.

En korrelationsmatris används för att analysera sambandet mellan olika variabler. I en regressionsanalys hjälper det att identifiera vilka variabler som har starkast samband med målet (i detta fall bilpriset). En hög korrelation (positiv eller negativ) kan indikera viktiga prediktorer för modellen.

5.2 Multipla linjära regressionen

är en klassisk statistisk metod som försöker modellera sambandet mellan en beroende variabel (bilens pris) och flera oberoende variabler (såsom årsmodell, miltal, motortyp och så vidare). Modellen bygger på antagandet att sambandet mellan variablerna är linjärt.

$$Y=\beta 0+\beta 1x1+\beta 2x2+\cdots+\beta kxk+\varepsilon$$

Där:

- Y är den beroende variabeln (det predikterade priset).
- $\beta 0$ är interceptet (värdet på Y när alla xi är 0).
- $\beta 1,\beta 2,\ldots,\beta k$ är regressionskoefficienterna som representerar förändringen i Y vid en enhets förändring i respektive xi.
- ε är feltermen (residualen) som fångar slumpmässiga variationer som modellen inte kan förklara. (simplilearn, u.d.)

5.3 XGBoost

(Extreme Gradient Boosting) är en typ av boosting-algoritm som använder gradientnedstigning. Liksom andra boosting-metoder, börjar gradientboosting med en svag lärandealgoritm för att göra förutsägelser. Det första beslutsträdet i gradientboosting kallas för basläraren. Därefter skapas nya träd på ett additivt sätt baserat på baslärarens misstag. Algoritmen beräknar sedan residualerna för varje träds förutsägelser för att avgöra hur långt modellen var från verkligheten. Residualer är skillnaden mellan modellens förutsagda och faktiska värden. Residualerna sammanställs sedan för att bedöma modellen med en förlustfunktion. (IBM, u.d.)

5.4 Utvärderingsmått

För att utvärdera modellernas prestanda har vi använt tre vanliga mått:

Root Mean Squared Error (RMSE)

RMSE visar hur mycket fel modellen gör i genomsnitt, där större fel får större påverkan. Ju lägre RMSE, desto bättre är modellen.

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

där:

- y_i är de faktiska värdena
- $\hat{y}_i$ är de predikterade värdena
- n är antalet observationer

Mean Absolute Error (MAE)

MAE ger ett mått på det genomsnittliga felet mellan de predikterade och faktiska priserna. Lägger lika mycket vikt på alla fel, oavsett om de är små eller stora.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

R^2 (Förklaringsgrad)

R^2 visar hur stor andel av variationen i priset som modellen kan förklara. Ett högre R^2 innebär att modellen gör ett bättre jobb med att förklara variationen i data.

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

6 Metod

I denna sektion beskriver vi ursprunget och förberedelserna av datan. Vi förklarar också hur vi valde lämpliga maskininlärningsmodeller, samt hur vi utvärderade deras prestanda. Vidare går vi igenom de experiment vi genomfört för att träna och testa modellerna.

6.1 Data

Datan som användes i projektet samlades in manuellt av tolv personer, där varje person ansvarade för att samla cirka 100 observationer. Totalt samlades information om 1205 begagnade Volkswagen-bilar, producerade mellan åren 2000 och 2022.
De variabler som registrerades inkluderade bland annat årsmodell, modellnamn, körsträcka, drivmedel, växellåda och pris.
Syftet med datainsamlingen var att skapa ett representativt urval av bilar för att kunna analysera vilka faktorer som påverkar bilpriset.

6.2 Databehandling och Rensning

Innan analysen påbörjades genomfördes en noggrann förbehandling av datan.
Stegen inkluderade

- Vi använde R för att rensa och strukturera datan..
- Borttagning av dubbletter.
- Hantering av saknade värden genom att ta bort eller fylla i med rimliga uppskattningar.
- Standardisering av textdata (exempelvis rättning av stavfel och borttagning av onödiga mellanslag).
- Identifiering och hantering av avvikande värden (outliers), där extremt höga eller låga priser och körsträckor analyserades och i vissa fall exkluderades för att inte snedvrida modellerna.

6.3 Modellering

Vi byggde två modeller för att prediktera bilpriser: Multipel Linjär Regression och XGBoost Regression.

6.3.1 Dataluppdelning

Datan delades i tre delar:

- Träningsdata (60%): Användes för att träna modellerna.
- Validationsdata (20%): Användes för modellvalidering.
- Testdata (20%): Användes för att utvärdera modellernas slutgiltiga prestanda.

Uppdelningsmetod:

- Först valdes 60% slumpmässigt till träningsdata.
- Resterande 40% delades upp i två lika stora delar (20% + 20%) för validering och testning.

6.3.2 Två regression modeller testades:

Modell 1: Multipel Linjär Regression

- Feature Selection:
    - Vi använde en stegvis variabelselektion (stepwise selection) för att välja de viktigaste prediktorerna för modellen.
- Modellträning:
    - Den linjära regressionen tränades på träningsdatan och justerades baserat på valideringsresultat.
- Utvärdering:
    - Modellen utvärderades på testdatan med metrik såsom RMSE, MAE och R^2.

Modell 2: XGBoost Regression

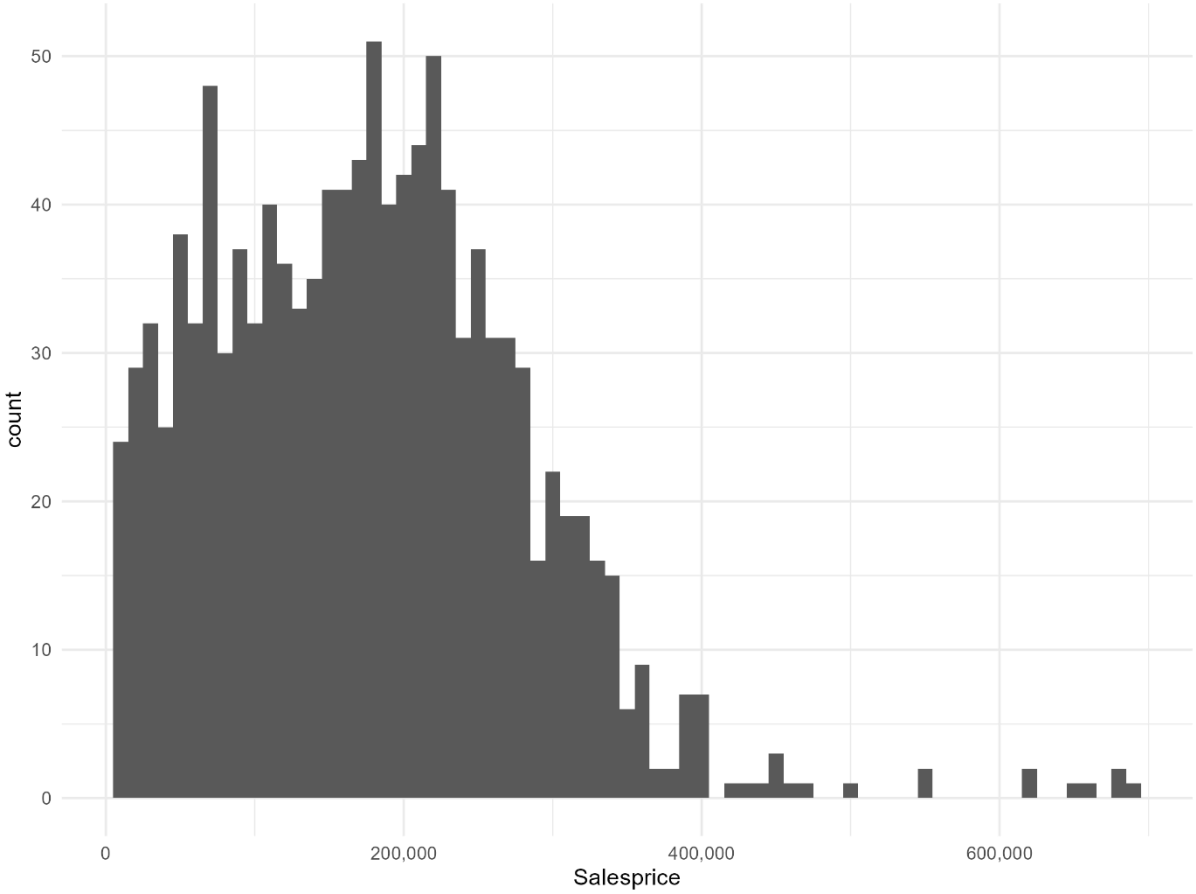
- Feature Selection:
    - Vi tränade först en fullständig XGBoost-modell.
    - Sedan använde vi xgb.importance() för att extrahera variabelviktigheter.
    - Därefter valde vi de topp 5 viktigaste prediktorerna för att bygga en enklare, mer fokuserad modell.
- "För mer detaljer om koden som användes för att träna XGBoost-modellen, Appendix A.

7 Resultat och Diskussion

7.1 Explorativ Dataanalys (EDA)

Vi började analysen med en utforskande dataanalys (EDA) för att förstå datans struktur och mönster. Syftet var att identifiera:

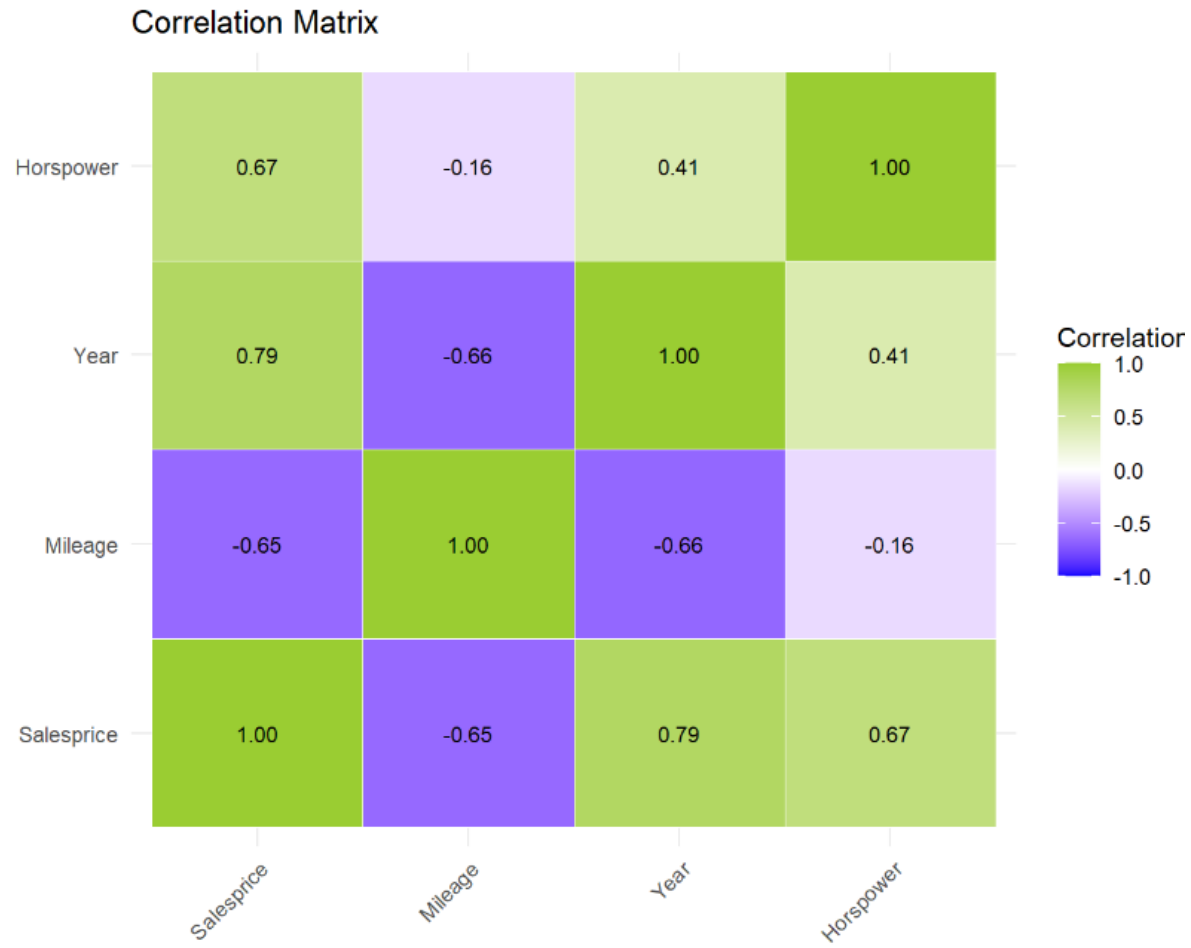
- Distributioner av variabler
- Korrelationsmönster mellan variabler
- Outliers och eventuella avvikande värden
- Behov av datatransformationer (exempelvis log-transformation)



Figur 3:sales price distribution after removing outliers

7.1.1 korrelationsmatris

För att bättre förstå sambanden mellan de numeriska variablerna i datan beräknades en korrelationsmatris.
Korrelationsmatrisen hjälper till att identifiera starka positiva eller negativa samband mellan variabler, vilket är viktigt både för feature selection och för att upptäcka multikollinearitet i modelleringssteget.



Figur 4:correlation matrix for numeric variables

7.1.2 Outliers

Vid den utforskande dataanalysen (EDA) identifierades tre tydliga outliers:

- Två observationer med extremt höga bilpriser jämfört med övriga datan.
  Vid vidare granskning visade det sig att en av dessa var en specialbeställd bil och den andra en rallybil, vilket motiverade deras ovanliga prisnivåer.
- En observation med ett extremt högt miltal som inte stämde överens med bilens angivna pris, vilket indikerade ett möjligt datainmatningsfel eller en ovanlig försäljningssituation.

För att säkerställa att modellerna inte skulle bli snedvridna av dessa extrema värden, togs dessa tre observationer bort från datamängden innan vidare analys och modellering.

7.2 Datatransformation och Feature Engineering

För att förbättra modellernas prestanda och hantera olika typer av variabler, utfördes följande transformationer:

- **Log-transformering** av målvariabeln *Salesprice* för att minska snedfördelningen och stabilisera variansen. Detta steg utfördes innan modellträningen.
- **Skapande av dummie-variabler** för kategoriska variabler (t.ex. bränsletyp, växellåda) så att modellerna kunde hantera dem korrekt. Dummifieringen genomfördes i modellbyggnadsfasen.

7.3 Modellutveckling och Utvärdering

7.3.1 Linear model summary

- Added row_id to train_data for tracking.
- Standardized numeric variables (Mileage, Year, Horspower) using training set statistics.
- Trained a full linear model (lm_train_full)
- Used step() for stepwise selection (lm_stepwise), direction = "both", trace = 0.
- Summary of lm_stepwise shows 26 selected variables, including Salesprice, Mileage, Year, Horspower, Seller type, Fuel type, Gearbox type, Cartype, and Modell.
- R^2 on training set: 0.9446, adjusted R^2: 0.9425.

Validation of linear model

- **Validation RMSE**: 27868.36
- **Validation MAE**: 21065.49
- **Validation R^2**: 0.9297

7.3.2 XGBoost Model Summary

• Prepared data for XGBoost by creating a DMatrix (dtrain) excluding the target variable Salesprice_log as features.
• Trained a full XGBoost model (xgb_model_full) with the objective reg:squarederror and 100 boosting rounds.
• Feature importance calculated using xgb.importance, showing key features: Salesprice, Year, Mileage, Horspower, row_id.
• Top 5 selected features based on importance:

- Salesprice
- Year
- row_id
- Mileage
- Horspower
  • RMSE on training set: 0.001164 after 100 iterations.

Validation of XGBoost Model

- Validation RMSE: 4630.503
- Validation MAE: 3740.058
- Validation R^2: 0.9981

7.3.2 Modelljämförelse

Efter att ha utvärderat prestandan hos både den **linjära modellen** och **XGBoost-modellen**, baseras beslutet att fortsätta med **XGBoost-modellen** på dess överlägsna prestandamått.

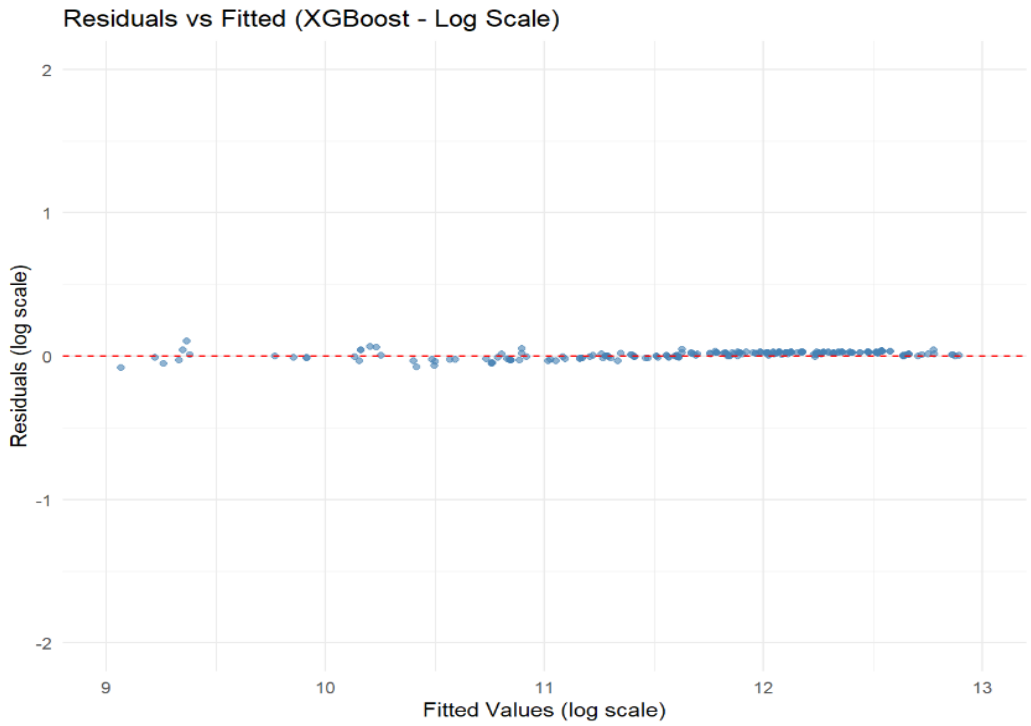
Den **linjära modellen** visade ett **R^2 på 0,9297** på valideringssetet, vilket innebär att cirka 93 % av variansen i målvariabeln (Salesprice) förklarades av de valda funktionerna. Däremot var dess **validerings-RMSE (27 868,36)** och **validerings-MAE (21 065,49)** relativt höga, vilket tyder på att den linjära modellens förutsägelser inte var lika precisa, särskilt för individuella observationer.

I kontrast levererade **XGBoost-modellen** ett **R^2 på 0,9981** på valideringssetet, vilket innebär en nästan perfekt anpassning till data. Detta betyder att XGBoost förklarade 99,81 % av variansen i målvariabeln. **Validerings-RMSE (4 630,50)** och **validerings-MAE (3 740,06)** är betydligt lägre, vilket tyder på att XGBoost-modellens förutsägelser är mycket mer precisa jämfört med den linjära modellen.

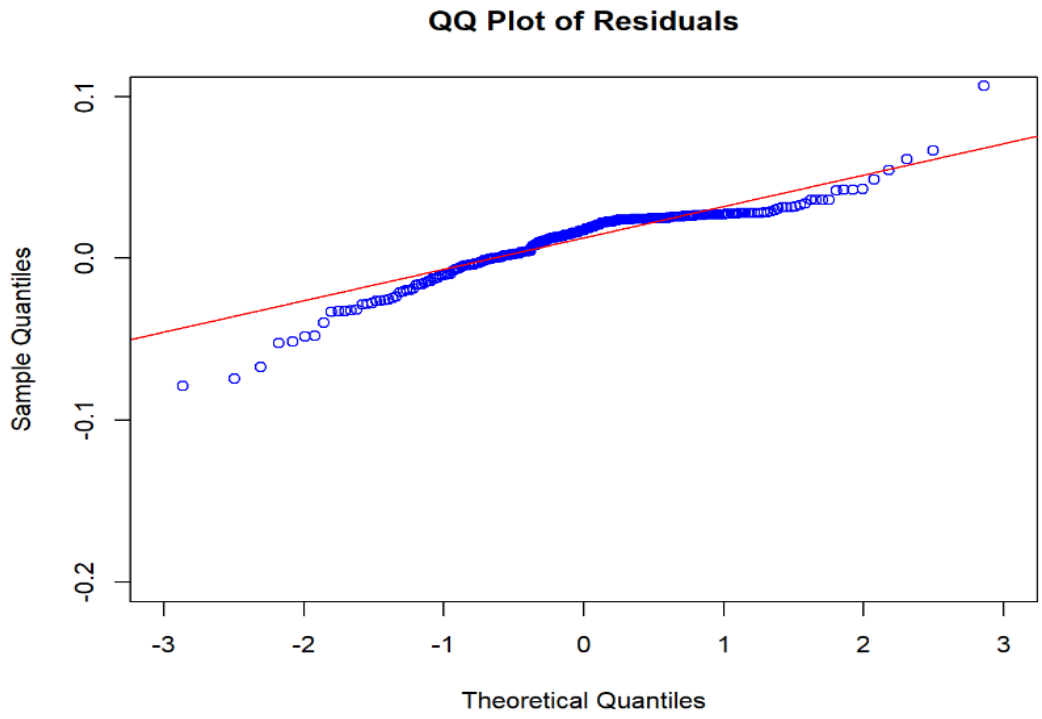
Jämförelse av Modellens Prestanda på Testdata

Metrik	Linjär Modell	XGBoost
RMSE	32 295,61	4 396,60
MAE	21 333,60	3 337,68
R^2	0,8999	0,9981

7.3.3 Visualisering



Figur 5:residual plot



Figur 6:QQplot

8 Slutsats

Efter att ha fått fantastiska resultat från XGBoost ville jag säkerställa att min modell skulle generalisera på helt osedda data. Därför samlade jag in 10 nya observationer från Blocket och förutspådde priserna med XGBoost. Här är resultaten:

Mileage	Row ID	Year	Horsepower	Actual Price	Predicted Price	Error %
21000	1501	2008	102	23000	22939.47	-0.26%
7088	1502	2021	151	284900	284824.50	-0.03%
9526	1503	2021	218	229800	229789.78	0.00%
10084	1504	2017	191	209800	209886.07	0.04%
15599	1505	2017	191	89800	88853.30	-1.05%
24041	1506	2005	175	40000	40103.52	0.26%
12577	1507	2020	150	208900	209071.79	0.08%
6300	1508	2014	105	129900	129780.22	-0.09%
19739	1509	2012	170	109900	109757.80	-0.13%
6340	1510	2018	111	178800	179057.43	0.14%

Vad detta betyder:

- Modellen har förutspått priserna för samtliga 10 nya observationer, och de förutspådda priserna är mycket nära de faktiska priserna.
- Eror Percent visar skillnaden mellan det förutspådda priset och det faktiska priset i procent. Den största skillnaden i procent var 2.44%, vilket innebär att modellen förutspådde priset för en bil med ett litet fel.
- För de flesta observationerna är felet mycket litet, vilket visar att modellen inte bara har presterat bra på tränings- och testdata utan också generaliserar väl till nya, osedda data.
- Detta bekräftar modellens robusthet och förmåga att göra exakta förutsägelser på verkliga data.

Sammanfattningsvis innebär detta att **XGBoost-modellen** inte bara fungerar bra på tränings- och testdata, utan också förutsäger priser på osedda observationer med mycket liten felmarginal, vilket gör den till ett pålitligt verktyg för att förutsäga begagnade bilpriser.

1. Hur väl kan multipel linjär regression och XGBoost förutsäga priser på begagnade bilar?

Både multipel linjär regression och XGBoost visade sig vara användbara för att förutsäga priser på begagnade bilar, men XGBoost överträffade linjär regression i prestanda.

- Multipel linjär regression gav ett R^2 på 0.9297 på valideringsuppsättningen, vilket innebär att modellen förklarade 93% av variansen i priset. Trots detta var felmåtten (RMSE och MAE) relativt höga, vilket antyder att modellen inte var lika exakt i sina individuella förutsägelser. Efter att ha tagit bort extrema avvikande observationer (outliers) förbättrades dock modellens prestanda markant.
- XGBoost presterade mycket bättre med ett R^2 på 0.9981 på valideringsuppsättningen, vilket innebär att modellen förklarade nästan 100% av variansen i data. Dess RMSE och MAE var också avsevärt lägre än linjär regression, vilket tyder på att XGBoost gav mycket mer precisa och pålitliga förutsägelser. När extrema outliers togs bort visade XGBoost en ännu starkare förmåga att generalisera och förutsäga priser korrekt.

2. Vilka variabler har störst inverkan på prissättningen av begagnade bilar?

- År (Year): Äldre bilar har vanligtvis ett lägre pris, medan nyare bilar tenderar att behålla ett högre värde. Detta var en av de mest betydelsefulla variablerna både i XGBoost och linjär regression.
- Körsträcka (Mileage): Färre körda kilometer innebär ett högre värde på bilen. Ju högre körsträcka, desto lägre pris, vilket var tydligt i båda modellerna.
- Hästsyrka (Horsepower): Bilar med högre hästkrafter tenderade att ha ett högre pris, vilket reflekterades i både linjär regression och XGBoost.
- Biltyp (Car type) och Bränsletyp (Fuel type): Dessa variabler visade också en viss påverkan på prissättningen, men inte lika stark som de andra faktorerna. Lyxigare bilmodeller och bränsleeffektivitet verkade också spela en roll.

XGBoost, genom sin förmåga att hantera icke-linjära samband och interaktioner mellan variabler, identifierade att andra faktorer som modellnamn och säljartyp också var viktiga för att förutsäga priser.

3. Hur påverkar extrema avvikande observationer (outliers) modellernas prestanda och prediktioner?

Extrema avvikande observationer (outliers) hade en betydande negativ inverkan på modellernas prestanda och förutsägelser. I början, när outliers var inkluderade i datamängden, presterade både den linjära regressionen och XGBoost dåligt:

- Multipel linjär regression var känslig för outliers, vilket resulterade i en högre RMSE och MAE. Det ledde till att modellen inte lyckades förutsäga priser korrekt, särskilt för bilar med mycket ovanliga eller extrema attribut.
- XGBoost var något mer robust mot outliers, men även denna modell visade försämrad prestanda när extrema värden var närvarande. XGBoost arbetar bra med större datamängder och komplexa relationer, men outliers kan fortfarande störa modellens förmåga att generalisera och göra precisa förutsägelser.

Framtida Möjligheter och Förbättringar

För framtiden finns flera områden där vi kan fortsätta att förbättra modellen och analysera ytterligare faktorer som kan påverka prissättningen av begagnade bilar:

1. Inkludera fler variabler: För att ytterligare förbättra modellens prediktiva förmåga kan fler variabler övervägas. Exempelvis kan detaljer om bilens skick (t.ex. repor, rost eller tidigare skador), servicehistorik, och geografisk plats ha en påverkan på priset som inte fångades i denna modell. Att utforska dessa ytterligare variabler kan ge en mer komplett och exakt prissättningsmodell.
2. Integrera mer uppdaterad och större dataset: För att säkerställa att modellen fortsätter att vara relevant och exakt kan det vara fördelaktigt att kontinuerligt uppdatera och integrera nya data om begagnade bilar från olika källor som exempelvis Blocket, Tradera eller andra bilsajter.

Genom att ta dessa faktorer i beaktning kan vi fortsätta att förbättra och anpassa modellen för att ge mer precisa och pålitliga prissättningar av begagnade bilar.

9 Teoretiska frågor

1. beskriv kortfattat vad en Quantile-Quantile (QQ) plot är.

Inom statistiken är ett Q–Q-diagram (quantile–quantile plot) ett sannolikhetsdiagram, en grafisk metod för att jämföra två sannolikhetsfördelningar genom att plotta deras kvantiler mot varandra. En punkt (x, y) i diagrammet motsvarar en av kvantilerna från den andra fördelningen (y-koordinaten) plottad mot samma kvantil från den första fördelningen (x-koordinaten). Om de två fördelningarna som jämförs är lika, kommer punkterna i Q–Q-diagrammet att ligga ungefär längs identitetslinjen y = x. Om fördelningarna är linjärt relaterade, kommer punkterna i Q–Q-diagrammet att ligga ungefär längs en rät linje.

2.Din kollega Karin frågar dig följande: *”Jag har hört att i Maskininlärning så är fokus på prediktioner medan man i statistisk regressionsanalys kan göra såväl prediktioner som statistisk inferens. Vad menas med det, kan du ge några exempel?”* Vad svarar du Karin?

Ja, du har helt rätt!
I maskininlärning är fokus främst på prediktion ,målet är att få så bra förutsägelser som möjligt på nya osedda data. Man bryr sig mindre om hur och varför modellen gör sina förutsägelser, och mer om hur bra den förutsäger. Exempel: en maskininlärningsmodell som förutspår huspriser behöver inte förklara exakt hur varje faktor påverkar priset, så länge den gissar rätt.

I statistisk regressionsanalys (som linjär regression) kan man både göra prediktioner och statistisk inferens. Inferens betyder att vi vill förstå och dra slutsatser om sambanden i datan. Till exempel: vi kan testa om ett samband är statistiskt signifikant, vi kan tolka koefficienter ,hur mycket ökar priset om huset blir 10 kvm större

3. Vad är skillnaden på ”konfidensintervall” och ”prediktionsintervall” för predikterade värden?

Konfidensintervall: För osäkerheten kring medelvärdet av prediktionerna.

Prediktionsintervall: För osäkerheten kring ett individuellt predikterat värde.

4.Den multipla linjära regressionsmodellen kan skrivas som: $Y = \beta 0 + \beta 1x1 + \beta 1x2 + \ldots + \beta pxp + \varepsilon$.

Hur tolkas beta parametrarna?

Y är den beroende variabeln (det vi försöker förutsäga) och $x1$, $x2$ och xp är de oberoende variablerna.

β tolkas på följande sätt:
β0 är interceptet, vilket betyder det predikterade värdet av YYY när alla oberoende variabler $x1$, $x2$ och xp är lika med noll.

βi representerar ändringen i den beroende variabeln Y när den oberoende variabeln xi ökar med en enhet, med alla andra oberoende variabler konstant.

Epsilon representerar den slumpmässiga felterm eller den del av variationen i Y som inte förklaras av modellen.

5. Din kollega Nils frågar dig följande: ”Stämmer det att man i statistisk regressionsmodellering inte behöver använda träning, validering och test set om man nyttjar mått såsom BIC? Vad är logiken bakom detta?” Vad svarar du?

BIC och AIC kan minska risken för överanpassning, men du behöver ändå överväga hur modellen presterar på nya data, vilket ibland gör det lämpligt att använda tränings- och testuppdelningar.I strikt statistisk regressionsmodellering är BIC och AIC en metod för att välja en bra modell, men testning på nytt data är fortfarande viktig för att säkerställa bra generalisering.

6.Förklara algoritmen nedan för ”Best subset selection”

Starta med nulllmodellen (utan prediktorer).

För varje k (antal prediktorer), testa alla möjliga modeller med exakt k prediktorer och välj den bästa modellen (den med lägst RSS eller högst R^2).

Jämför alla modeller (från M0 till Mp) och välj den modell som har bäst prediktionsförmåga baserat på mått som AIC, BIC, justerat R^2 eller cross-validation.

7. Ett citat från statistikern George Box är: “All models are wrong, some are useful.” Förklara vad som menas med det citatet.

Det handlar om modellernas begränsningar och deras praktiska värde. Alla modeller har sina brister och kan inte exakt representera verkligheten, men så länge de ger användbara och tillförlitliga resultat för det syfte de är avsedda för, är de fortfarande värdefulla.

10 Självutvärdering

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

En av de största utmaningarna under arbetet var datarensningen i R, särskilt när det gällde att hantera svenska tecken vid kodning och när jag knöt samman rapporten till HTML. Jag stötte på flera problem med teckenkodning som orsakade att svenska bokstäver som å, ä och ö inte visades korrekt i rapporten.

Lösning: För att lösa detta problem justerade jag teckenkodningen i R och säkerställde att rätt inställningar användes både vid importen av data och när jag genererade HTML-rapporten. Detta hjälpte till att säkerställa att alla tecken hanterades korrekt.

Trots att vi själva samlade in data och kom överens om hur det skulle göras, uppstod det ändå många fel i datauppsättningen. Detta gav mig en stor förståelse för hur verklig data ofta ser ut, med sina många brister och problem. Jag lärde mig mycket av den här processen och insåg hur viktigt det är att noggrant hantera och förbereda data innan analys.

En annan utmaning var hanteringen av extrema avvikelser (outliers) i datasetet. Dessa påverkade modellernas prestanda negativt, vilket ledde till mindre tillförlitliga förutsägelser.

Lösning: Jag identifierade och tog bort dessa outliers från datan, vilket förbättrade modellernas resultat avsevärt och gav mer exakta prediktioner.

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

Jag anser att jag bör få betyget VG eftersom jag har genomfört projektet med noggrant arbete och har lärt mig och tillämpat nya tekniker, trots de utmaningar som uppstod.

3.Något du vill lyfta fram till Antonio?

Jag vill verkligen tacka dig för allt stöd och all kunskap du har delat med dig av under kursen. Det har varit en värdefull lärandeupplevelse, och jag har fått med mig mycket. Din feedback förra gången hjälpte mig verkligen, och jag skulle vara tacksam om du kan ge mig feedback den här gången också. Det skulle hjälpa mig att fortsätta utvecklas inför kommande kurser. Stort tack igen för all hjälp!

11 Appendix A

```
library(xgboost)

library(dplyr)

set.seed(123)

# Train XGBoost model

dtrain <- xgb.DMatrix(data = as.matrix(train_data[, -which(names(train_data
) == "Salesprice_log")]), label = train_data$Salesprice_log)

xgb_model <- xgboost(data = dtrain, objective = "reg:squarederror", nrounds
= 100)
```

```
# Feature Importance for XGBoost

importance_xgb <- xgb.importance(feature_names = colnames(train_data[, -whi
ch(names(train_data) == "Salesprice_log")]), model = xgb_model)

print(importance_xgb)
```

```
#select top features 5 based on importance

top_features_xgb <- importance_xgb$Feature[1:5]

train_xgb_selected <- train_data[, c(top_features_xgb, "Salesprice_log")]


# Re-train XGBoost model with selected features

dtrain_selected <- xgb.DMatrix(data = as.matrix(train_xgb_selected[, -which
(names(train_xgb_selected) == "Salesprice_log")]), label = train_xgb_select
ed$Salesprice_log)

xgb_model_selected <- xgboost(data = dtrain_selected, objective = "reg:squa
rederror", nrounds = 100)
```

12 Källförteckning

(u.d.). Hämtat från Blocket: https://www.blocket.se/annonser/hela_sverige?q=bilar

IBM. (u.d.). Hämtat från https://www.ibm.com/think/topics/xgboost

simplilearn. (u.d.). Hämtat från https://www.simplilearn.com/what-is-multiple-linear-regression-in-machine-learning-article