Prisförutsägelser på Begagnad Bilmarknad



Parsan Amani

EC Utbildning

2024-04

# 

# Abstract

In this project, we built a Support Vector Machines (SVM) regression model to predict car prices using data from Blocket website and additional economic data from Statistics Sweden (SCB). The model combined specific car details like brand, year, and mileage with broader economic factors, and was fine-tuned through thorough preprocessing and optimization. It achieved high accuracy, surpassing simpler models and demonstrating the value of machine learning in enhancing price prediction in the car market.

# Förkortningar och Begrepp

MSE : Mean Squared Error

MAE: Mean Absolute Error

SVM: Support Vector Machine

**Skapas automatiskt i Word genom att gå till Referenser > Innehållsförteckning.**

Table of Contents

[Abstract 2](#_Toc165059362)

[Förkortningar och Begrepp 3](#_Toc165059363)

[1 Inledning 1](#_Toc165059364)

[2 Teori 2](#_Toc165059365)

[2.1 Regressionsanalys 2](#_Toc165059366)

[2.1.1 Enkel Linjär Regression 2](#_Toc165059367)

[2.1.2 Multipel Linjär Regression 2](#_Toc165059368)

[3 Metod 3](#_Toc165059370)

[4 Resultat och Diskussion 4](#_Toc165059371)

[5 Slutsatser 6](#_Toc165059372)

[6 Teoretiska frågor 7](#_Toc165059373)

[7 Självutvärdering 9](#_Toc165059374)

[Appendix A 10](#_Toc165059375)

[Källförteckning 11](#_Toc165059376)

# Inledning

Den globala bilmarknaden genomgår stora förändringar på grund av digitalisering och ökad efterfrågan på hållbarhet, vilket påverkar både tillverkare och konsumenter. I Sverige är noggrann prissättning på bilar viktig, inte bara som transportmedel utan även som ekonomiska och miljömässiga investeringar. Med fler än en miljon nya personbilar i trafik de senaste 20 åren är precisa och effektiva prissättningsmodeller essentiella för att hjälpa både köpare och säljare.

Blocket.se erbjuder en omfattande datamängd som reflekterar bilprisernas variation och fungerar som en robust bas för att utveckla prediktiva modeller. Tillsammans med makroekonomiska data från Statistiska centralbyrån (SCB) kan dessa modeller förutse bilpriser med hög precision, vilket visar maskininlärningens potential i att förbättra beslutsprocesser inom bilhandeln.

Syftet med denna rapport är att skapa en kraftfull regressionsmodell som använder maskininlärningsmetoder, specifikt Support Vector Machines (SVM), för att förutsäga priser på begagnade bilar i Sverige. Modellen kommer att kombinera data från Blocket och Statistiska centralbyrån (SCB) för att tydligt identifiera de faktorer som mest påverkar bilpriserna, för att uppfylla syftet så kommer följande frågeställning att besvaras:

Hur noggrant kan SVM-modellen, som använder data från både Blocket och SCB, förutsäga priser på begagnade bilar i Sverige med särskilda ”features” som märke, år, bränsle, körsträcka och växellåda?

# Teori

## Regressionsanalys

Regressionanalys är en statistisk metod för prediktiv modellering som analyserar sambandet mellan en beroende variabel och en eller flera förklarande oberoende variabel. Målet med denna analys är att skapa en modell som kan förutsäga den beroende variabelns värde baserat på värdena av de oberoende variablerna. Denna teknik tillämpas inom flera områden som ekonomi, biologi, marknadsföring och samhällsvetenskap för att förutsäga utfall och identifiera vilka faktorer som spelar en avgörande roll i olika fenomen.

### Enkel Linjär Regression

Enkel linjär regression är en grundläggande statistisk teknik som använder en oberoende variabel X för att förutsäga en beroende variabel Y. Detta antar en approximativt linjär relation mellan X och Y, vilket matematiskt kan beskrivas med modellen:

därär skärningspunkten och är lutningskoefficienten, vilka tillsammans utgör modellens koefficienter eller parametrar. Denna modell är särskilt användbar för enkel analys, som att förutsäga försäljning baserat på en enskild marknadsföringskostnad, såsom TV-reklam.

#### Exempel på tillämpning:

*Om X representerar TV-reklambudget (i tusentals kronor) och Y representerar försäljning (i tusen enheter), kan en regression av Y på X hjälpa till att förutsäga försäljning baserat på hur mycket som spenderas på TV-reklam*

### Multipel Linjär Regression

### För situationer där flera oberoende variabler (prediktorer) påverkar den beroende variabeln, är multipel linjär regression en mer tillämplig metod. Denna teknik utvidgar den enkla linjära regressionen genom att inkludera flera prediktorer, var och en med en separat lutningskoefficient: där *Xj*​ representerar den j prediktorn och *βj* kvantifierar sambandet mellan den prediktorn och responsen Y

#### Tillämpning på orojekt:

#### I detta projekt, där målet är att förutsäga priser på bilar från Blocket med data som även innefattar externa makroekonomiska faktorer från SCB, tillämpas multipel linjär regression. Detta möjliggör en omfattande analys där flera variabler, som årsmodell, miltal, bränsletyp, samt ekonomiska indikatorer, simultant kan bedömas för deras påverkan på bilpriserna.

# Metod

**Datainsamling:**

Det finns sex personer i gruppen som vi börjad jobba med. Ming, Nike, Mustafa, Mohammad och jag. Vi hade två/tre online Teams möte tillsammans och hade bra diskutioner om hur ska vi bestämma att jobba med datan. Den här tanken att arbeta i ett grupparbete var så produktiv eftersom man kan spara tiden. Till exempel Nike är bra me Excel och kunde lära mige lite när vi ville rensa vissa datan på Excel och Mustafa är bra med att hantera datan och köra olika metoder snabbt på R. Så jag också kommer med tanken att köra Multiple Linear Regression i början som vi hade fyra features som med hjälp av dom skulle vi förutsäga billpriser igenom datan.

Vi samlade data om begagnade bilar från Blocket.se, Sveriges största online-marknadsplats. Vi valde annonser från olika regioner i Sverige för att få en mångsidig bild av bilmarknaden. Data för olika bilmodeller, årsmodeller och priser samlades in. Vi extraherade viktig information från varje annons inklusive märke, modell, årsmodell, pris, miltal, bränsletyp, växellåda och datum för annonspublicering. Vi rensade datan för fel och ofullständigheter, som felaktiga format och dubbletter, Vi valde multipel linjär regression och i nästa steg som vi skulle jobba individuelt jag valde Support Vector Machine (SVM) baserat på deras förmåga att hantera olika typer av samband. Modellerna tränades och finjusterades med en uppdelning av datat, där 80% användes för träning. Jag mätte modellens effektivitet med statistiska mått som MSE och MAE.

Residualanalyser hjälpte mig att identifiera eventuella problem med modellen, som icke-linjaritet. Detta tillvägagångssätt säkerställer en noggrann metod för att utveckla en effektiv modell för att förutsäga priser på begagnade bilar i Sverige.

# Resultat och Diskussion

Jämfört med den linjära regressionsmodellen visade SVM-modellen en förbättring i både MSE och MAE, vilket tyder på att SVM bättre kunde hantera de icke-linjära sambanden mellan bilens egenskaper och dess pris. Tabellen nedan sammanfattar resultaten från båda modellerna:

|  |  |
| --- | --- |
| **MSE för olika modeller** | |
| Enkel Linjär Regression | 12201924743.0367 |
| SVM | 6294144479.09095 |

Tabell 1: Root Mean Squared Error (RMSE) för de fyra valda modellerna.

Detta MSE-värde för en enkel linjär regressionsmodell visar att den genomsnittliga kvadrerade skillnaden mellan de förutsedda och faktiska värdena är ganska stor, vilket antyder att modellen kanske inte passar data särskilt väl. Så jag bestämt att köra med SVM modell. MSE-värdet för SVM-modellen är avsevärt lägre än för den enkla linjära regressionen. Detta visar att SVM-modellen passar bättre till datan, med förutsägelser som ligger närmare de faktiska observerade värdena

Trots SVM-modellens relativt bättre prestanda, finns det fortfarande utrymme för förbättringar. Begränsningar i både modeller inkluderar potentialen för överanpassning, särskilt när antalet prediktorer är stort i förhållande till antalet observationer. Framtida forskning kan inrikta sig på att inkludera fler datapunkter, implementera tekniker för dimensionssänkning som huvudkomponentanalys (PCA), eller utforska mer avancerade ensemblemetoder.

Jag har kollat på SCB att se hur datan är att köpa bil i den senaste decenniet i Västra Götaland Region och det är resultatet med a graf som visar.

A graph with a line

Description automatically generated

En ökning av antalet bilar i en region kan indikera ökad efterfrågan, vilket kan leda till högre bilpriser. Omvänt kan en stabilisering eller minskning i antalet bilar tyda på en mättnad i marknaden, vilket potentiellt kan sänka priserna. Och det är bara en del som jag granulera som det är viktig i min analys.

Resultaten från denna studie erbjuder insikter som kan hjälpa både köpare och säljare på den begagnade bilmarknaden att fatta mer informerade beslut. Bilhandlare kan använda dessa modeller för att sätta mer konkurrenskraftiga priser baserade på detaljerade analyser av marknadsdata, medan köpare kan använda dem för att förhandla bättre priser.

# Slutsatser

SVM-modellen, som kombinerar data från både Blocket och SCB, har visat sig vara effektiv i att förutsäga priser på begagnade bilar i Sverige. Genom att använda egenskaper som märke, årsmodell, bränsletyp, körsträcka och växellåda, har modellen levererat precisa prisförutsägelser. Dess precision har verifierats genom användning av medelkvadratiskt fel (MSE) och medelabsolutavvikelse (MAE) som mätvärden, vilket bekräftar modellens förmåga att pålitligt förutsäga bilpriser utifrån de angivna attributen.

# Teoretiska frågor

1. En QQ-plot är ett diagram som används för att jämföra två sannolikhetsfördelningar. Genom att rita upp kvantilerna från två olika fördelningar mot varandra, kan man se om de liknar varandra. Om punkterna på QQ-ploten bildar en rak linje, betyder det att de jämförda fördelningarna är lika.

Man använder ofta QQ-plots för att se om data stämmer överens med en förväntad fördelning, som till exempel en normalfördelning. X-axeln på diagrammet visar kvantilerna från den teoretiska fördelningen och Y-axeln visar kvantilerna från det faktiska datat. En rak linje i diagrammet tyder på att datat matchar den teoretiska fördelningen bra. Om linjen inte är rak, visar det att datat avviker från den förväntade fördelningen, vilket kan vara ett tecken på att datat är skevt eller har andra ovanliga egenskaper. QQ-plots är viktiga eftersom de hjälper till att kontrollera om datat uppfyller vissa antaganden, vilket är nödvändigt för många statistiska tester och modeller.

1. Inom maskininlärning är det främsta målet att använda befintliga data för att göra förutsägelser eller klassificera information. Ofta är det mer betydelsefullt att en modell kan förutsäga framtida data än att exakt förstå varför eller hur den fungerar. Regressionsanalys används inte bara för att göra sådana förutsägelser, utan även för statistisk analys för att förstå sambanden mellan olika variabler. Detta gör det möjligt att bedöma och tolka effekten som olika oberoende variabler har på en beroende variabel.
2. Prediktionsintervall koncentrerar sig på att förutse kommande händelser, medan konfidensintervall hanterar tidigare eller nuvarande händelser. Konfidensintervall är värdefulla för att uppskatta parametrar med inkludering av osäkerhetsmarginaler. I motsats till detta används prediktionsintervall för att förutspå framtida observationer, vilka omfattar både osäkerheten i skattningen och den inneboende variationen. Dessa intervall är vidare och tillhandahåller en mer fullständig översikt som är praktisk för att förutsäga enskilda händelser.

1. Interceptet epresenterar det förväntade värdet av den beroende variabeln Y när värdena på alla oberoende variabler är noll. Även om detta scenario ofta inte har en praktisk betydelse, är interceptet kritiskt för att justera modellen korrekt. Slope-koefficienterna , beskriver hur den beroende variabeln Y förändras när en av de oberoende variablerna *Xi* ökar med en enhet, medan de övriga oberoende variablerna hålls konstanta; varje koefficient (där ***i*** sträcker sig från 1 till ***p***) är avgörande för att förstå relationen mellan varje specifik oberoende variabel och den beroende variabeln.
2. Det stämmer när vi bygger statistiska modeller, som i maskininlärning, delar vi ofta upp datan i tre grupper: träning, validering och testning. Detta gör vi för att se till att modellen inte bara fungerar bra på datan den har övat på utan att den också ska klara av att hantera riktiga situationer men det finns också ett verktyg som kallas BIC, vilket hjälper till att jämföra olika modeller om vi har flera att välja mellan. BIC bedömer modeller baserat på datan vi redan har, och tar inte med ny, osedd data. Därför är det viktigt att ha ett testset, alltså en grupp data som modellen inte har sett under träningen eller valideringen. Det är för att vi verkligen ska kunna testa hur bra modellen fungerar. Det här är extra viktigt när vi verkligen behöver kunna lita på vad modellen säger.
3. Algoritmen syftar till att hitta den bästa kombinationen av prediktorer för en modell genom att testa alla möjliga kombinationer och sedan välja den modell som balanserar mellan att passa data väl låg RSS, hög (som det tolkas som *bästa*) och att undvika överanpassning; M0 är den enklaste modellen som inte använder några prediktorer. Den förutspår helt enkelt medelvärdet av observationerna. För varje antal prediktorer 𝑘 från 1 till 𝑝(totala antalet prediktorer):Passa alla möjliga modeller som använder exakt 𝑘 prediktorer. Det finns (𝑝𝑘) möjliga modeller att överväga för varje 𝑘. Välj den bästa modellen för varje grupp baserat på minsta RSS (summan av kvadraterna av residualerna) eller största . Den här modellen kallas 𝑀𝑘. Eftersom RSS minskar och ökar monotoniskt med antalet prediktorer, behöver man en mer balanserad metod för att välja den bästa modellen. En modell med fler prediktorer kommer alltid att ha lägre RSS och högre , men detta kan leda till överanpassning. För att välja den bästa modellen använder vi antingen statistisk mått som BIC / AIC eller justera .
4. Det betyder att ingen statistisk modell är en perfekt spegling av verkligheten. Alla modeller är simplifieringar och inkluderar antaganden som inte fullständigt matchar den komplicerade och oförutsägbara naturen hos den faktiska världen. Varje modell lämnar något ute, vare sig det gäller data eller de processer den försöker återskapa. Det kan bero på att den data som finns tillgänglig är begränsad eller på de simplifieringar som görs för att göra problemet mer hanterbart. Det innebär även att vi måste vara medvetna om och kritiska mot modellernas begränsningar och inte förvänta oss att de ska förklara eller förutsäga allt heltäckande. Det påminner oss om att vi alltid bör vara kritiska mot de modeller vi använder, ständigt ifrågasätta dem och förbättra dem snarare än att se dem som definitiva lösningar. Det understryker också vikten av att välja rätt modell för det specifika syftet. En enklare modell kan vara tillräcklig för vissa uppgifter, medan mer komplexa modeller kan behövas för andra.

# Självutvärdering

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

Att arbeta med R var den första utmaningen, men det visade sig vara enklare jämfört med andra programmeringsverktyg vi använt hittills. Det var även fördelaktigt att använda R efter att ha lärt oss om maskininlärning, vilket hjälpte till att öka vår förståelse.

1. Vilket betyg du anser att du skall ha och varför.   
   Jag har jobbat med denna modell och känner att jag kan klara av det och bli godkänd, eftersom jag inte har lagt ner arbete på API för att nå VG.
2. Något du vill lyfta fram till Antonio? Jag vill bara uttrycka min tacksamhet till Antonio, som är en hjälpsam och empatisk person. Det finns inget mer att tillägga här.

# Appendix A

# Källförteckning

James, G., Witten, D., Hastie, T., & Tibshirani, R. (år). An Introduction to Statistical Learning with Applications in R. Förlag.

<https://www.youtube.com/watch?v=X9_ISJ0YpGw&t=290s>