Prediktiv modellering för begagnade fordon.

MOHAMAD ALMAZRLI

EC Utbildning

R_Programering

202404

Innehåll

1. Abstract

This study develops predictive models for used vehicles, Employing machine learning techniques, particularly random forest regresssion, n. The models explore the influence of brand, year, mileage, fuel type, and gearbox on prices. A diverse dataset ensures model robustness. Rigorous evaluation and cross-validation identify optimal configurations and hyperparameters, enhancing accuracy and reliability. Discussion includes data collection, preprocessing, and model interpretation, providing practical guidelines for future research and industry applications.

2 Inledning

I den här rapporten presenterar vi en analys av en dataset som samlats in från en Excel-filen , där vi arbetat med data i Excel-filen för att genomföra projektet. Syftet med denna studie är att utforska tillämpningen av random forest regression inom prediktiv analys, genom att utnyttja korsvalideringstekniker för att säkerställa robusthet och generalisering. Metodiken omfattar förbehandling av data, val och träning av modeller, korsvalidering och presentation av resultat. Genom en rigorös analysprocess syftar vi till att avslöja mönster, trender och relationer inom data, vilket ger värdefulla insikter för beslutsfattande. Genom att dokumentera vår metod och våra resultat bidrar vi till den växande kunskapsbasen inom dataanalys och maskininlärning, med fokus på reproducerbarhet och transparens.

Omfattning och begränsningar

Trots den enorma potentialen hos online marknadsplatser som Blocket är det viktigt att erkänna den inneboende komplexiteten och begränsningarna kring förutsägelsen av priserna på begagnade bilar. Utöver det annonserade priset ligger en myriad av faktorer som påverkar det slutliga transaktionsvärdet, inklusive körsträcka, servicehistorik, varumärkesrykte och säljarens motivation. Dessutom, även om denrandom forest regression valdes för dess robusta prediktiva prestanda, kan dess effektivitet begränsas av den relativt lilla datasats som är tillgänglig för analys. Med andra ord, det finns komplexa variabler och begränsningar som måste beaktas vid försök att förutsäga begagnade bilpriser.

Metodöversikt

Den valda metoden fokuserar på att använda en random forest regression , känd för sin förmåga att upptäcka komplexa mönster och samband i data. Genom att samla in och kombinera flera databasfiler med bilinformation skapades omfattande dataset, som sedan bearbetades och analyserades vidare i Excel-filer och R-programmering. Målet är att utveckla en prediktiv modell för att noggrant uppskatta priserna på begagnade bilar.

3 Teori

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

Svar. En Quantile-Quantile (QQ) plot är en grafisk metod för att jämföra fördelningar genom att placera kvantiler från en dataset mot motsvarande kvantiler från en teoretisk fördelning.

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?

Svar. I maskininlärning handlar det främst om att göra prediktioner medan statistisk regressionsanalys inkluderar både prediktioner och möjligheten till statistisk inferens, vilket innebär att man kan dra slutsatser om samband mellan variabler baserat på data.

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

Svar. Konfidensintervall ger en uppskattning av noggrannheten för en populationsparameter, medan prediktionsintervall ger en uppskattning av noggrannheten för individuella förutsägelser.

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

Låt oss överväga ett exempel där β1 = 2,5 i en multipellinjär regressionsmodell som försöker förutsäga studenters provresultat (Y) baserat på antalet timmar de studerade (X1) Om β1 = 2,5 betyder det att för varje ytterligare timme en student studerar (X1) deras provresultat (Y) förväntas öka med 2,5 poäng, förutsatt att alla andra faktorer förblir desamma. På liknande sätt, om β1 var negativ, skulle det innebära att för varje ytterligare studerad timme skulle undersökningen minska med det absoluta värdet av β1.

Jag skulle säga till Hassan att även om BIC kan hjälpa till med modellval, ersätter det inte behovet av utbildning, validering och testset. Dessa är avgörande för att säkerställa modellens generaliserbarhet och prestanda på ny data.

1. Initial modell utan prediktorer, förutsäger genomsnittligt utfall och fastställer baslinje för jämförelse.

2. För varje k från 1 till p, passa alla möjliga modeller som innehåller kk-prediktorer. Välj sedan modellen med den minsta restsumman av kvadrater (RSS) eller den största bestämningskoefficienten (R2). Denna iterativa process utforskar alla kombinationer av prediktorer och identifierar den bästa modellen för varje antal prediktorer k.

3. Algoritmen 'Bästa delmängdsval' fungerar så här: den provar olika kombinationer av prediktorer för att bygga modeller och väljer sedan den bästa genom att kontrollera hur väl den förutsäger ny data. Den använder saker som valideringsuppsättningsfel, Cp (AIC), BIC eller justerad R^2 för att avgöra vilken modell som är mest exakt utan att vara för komplicerad.

5. Din kollega Hassan 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 Hassan?

Svar.

I statistisk regressionsmodellering kan BIC användas för att välja den bästa modellen utan att behöva separata tränings-, validerings- och testset. Det beror på att BIC tar hänsyn till både modellens passning och dess komplexitet, vilket hjälper till att undvika överanpassning. Trots detta kan det fortfarande vara användbart att validera modellen mot oberoende data för att säkerställa dess generaliseringsförmåga.

6. Förklara algoritmen nedan för ”Best subset selection”

Svar. Best subset selection" är en regressionsanalysalgoritm som genererar alla möjliga kombinationer av variabler och utvärderar dem för att välja den bästa modellen baserat på ett utvärderingskriterium som AIC eller BIC. Målet är att hitta den optimala uppsättningen variabler som balanserar modellens passform till data med dess komplexitet.

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

Svar. Citatet betyder att även om modeller inte visar allt om det verkliga livet, kan de fortfarande vara praktiska.'

4. Metod

I det här avsnittet beskriver vi metodiken som används i vår dataanalysprocess, och beskriver de steg som vidtagits för att förbehandla data, välja och träna random forest regression, utföra korsvalidering och presentera resultaten.

4.1 Dataförbehandling:

Informationen om bilar samlades in från olika källor, vilka innehöll detaljer som växellåda, märke, årsmodell och pris. Därefter integrerades all data från dessa källor till en enda Excel-fil för vidare analys. Under bearbetningen av datan, rensades oönskade uppgifter bort, såsom priser för lyxbilar som Ferraris som ligger över miljonen, och också fejkdata elimineras för att säkerställa att endast pålitlig information används.

4.2 Val av modell och Training:

Random forest valdes för sin skalbarhet och bekantskap, vilket gör att den effektivt kan hantera ett brett spektrum av modeller.

Inledningsvis tränades modellen på 20% av datasetet, vilket resulterade i låg noggrannhet. Därefter förbättrades noggrannheten genom att träna på 30% av datasetet utan att kompromissa med modellens syfte.

4.3 Cross-Validation:

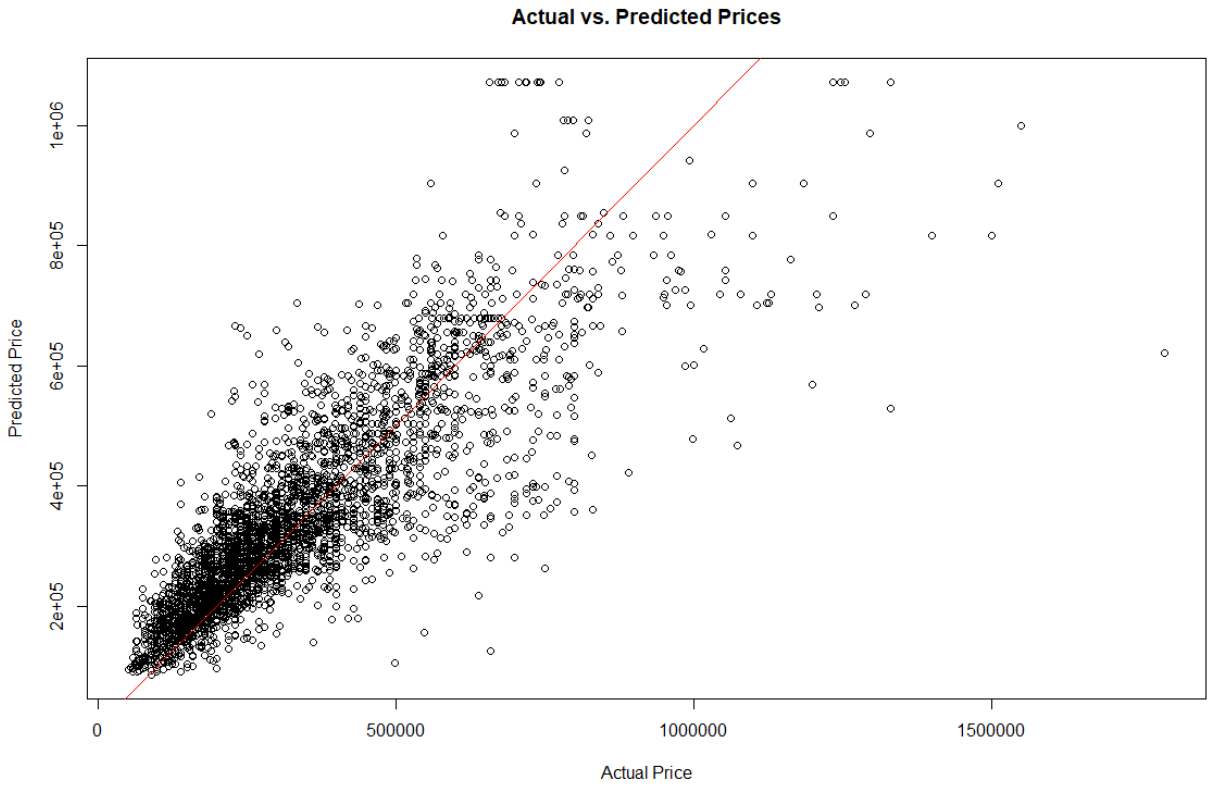
Genom att använda 10-faldig korsvalidering utvärderades modellens prestanda med hjälp av trainControl-funktionen i R's caret-paket.

Korsvalideringsresultaten visade varierande prestandamått över olika justeringsparametrar (mtry).

Den optimala modellen, med mtry = 22, valdes baserat på det minsta roten ur medelkvadratfelet (RMSE)-värdet.

4.4 Results:

Den random forest model tränades på ungefär 7800 prover.



Korsvaliderad omsampling över 10 gånger avslöjade följande prestandamått:

RMSE: 77 628,41

R-kvadrat: 0,7315

Genomsnittligt absolut fel (MAE): 54 427,00

4.5 Implementeringsdetaljer:

Analysen utfördes med R programmeringsspråk, med följande paket installerade: readxl, randomForest, caret, ggplot2, leaps, broom och MASS.

Nyckelfunktioner från caret-paketet, såsom trainControl och train, underlättade modellträning och utvärdering.

4.6 Reproduktion:

Nedan kommer ett kodexempel för framtida referens, för den som vill förbättra eller arbeta vidare med koden. Det tillhandahålls för att säkerställa att koden förblir korrekt och kan förbättras med tiden för att möta eventuella behov eller förbättringar.

5. Slutsatser

I den här studien använde vi random forest model tillsammans med korsvalideringstekniker för att analysera en datauppsättning från exel file som innehåller priser på begagnade bilar. Metodiken innebar att förbearbeta data, välja och träna modellen, utföra korsvalidering och presentera resultaten. Här är de viktigaste slutsatserna från analysen:

5.1Modellprestanda:

Den random forest model visade lovande prestanda, med en optimal konfiguration av mtry = 22 vilket resulterade i ett RMSE-värde på 77 628,41, R-kvadrat av 0,7315 och MAE på 54 427,00. Dessa metriker indikerar modellens förmåga att noggrant förutsäga målvariabeln.

5.2 Korsvalidering :

Användningen av 10-faldig korsvalidering gjorde det möjligt för oss att bedöma modellens robusthet och generaliseringsförmåga. Variabiliteten i prestandamått över olika inställningsparametrar understryker vikten av parameteroptimering för att uppnå optimala resultat.

5.3 Reproducerbarhet och transparens:

För att säkerställa reproducerbarheten och transparensen av vår analys tillhandahölls omfattande dokumentation och kodkommentarer, vilket gör det möjligt för framtida forskare att replikera vår metod med lätthet. Koden finns i bilagan.

5.4 Utmaningar och framtida riktningar:

Även om vår analys gav lovande resultat, möttes utmaningar under dataextraktionsprocessen, vilket underströk behovet av mer sofistikerade webbskrapningstekniker. Framtida forskning kan utforska alternativa datainsamlingsmetoder eller förfina befintliga tekniker för att övervinna dessa utmaningar.

Sammanfattningsvis visar vår studie effektiviteten av slumpmässiga skogsmodeller i prediktiv analys och betonar vikten av rigorös metodik, inklusive dataförbearbetning och korsvalidering, för att uppnå tillförlitliga resultat. Genom att dokumentera vår process och fynd bidrar vi till kunskapsmassan inom dataanalys och ger en grund för vidare forskning inom detta område.

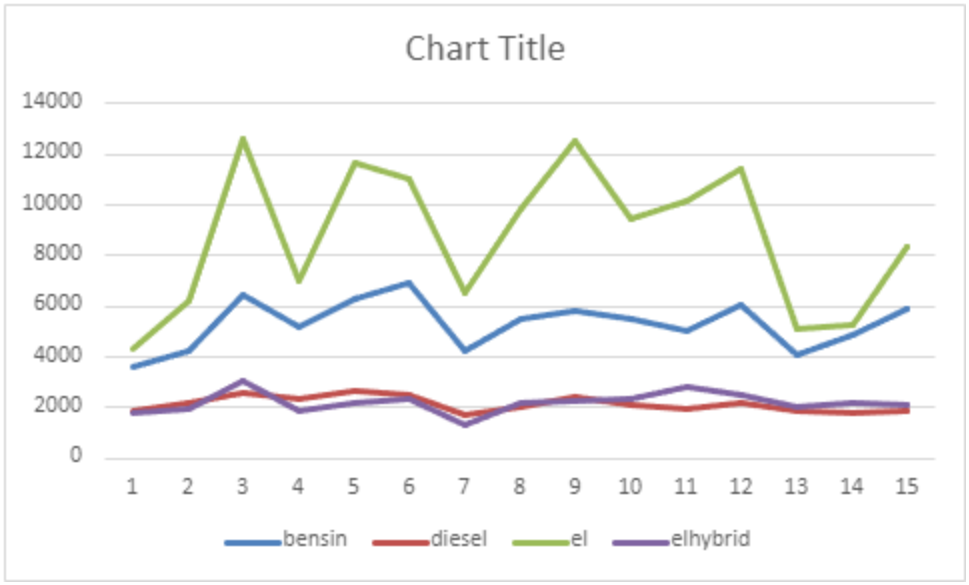
5.5 Utnyttjande av Extern Data för Utvecklingen av en Begagnad Bilplattform

I dagens samhälle, där hållbarhet och ekonomiskt ansvar blir allt viktigare, har efterfrågan på begagnade bilar ökat markant. Att förstå denna trend och de faktorer som påverkar begagnatbilsmarknaden kräver en analys av extern data för att skapa en effektiv och konkurrenskraftig plattform för begagnade bilköp.

Enligt Statistiska Centralbyrån (SCB) har försäljningen av begagnade bilar i Sverige ökat stadigt de senaste åren. Detta indikerar en växande preferens bland konsumenter för att köpa begagnade fordon, antingen av ekonomiska skäl eller med tanke på hållbarhet och miljöpåverkan.

Genom att nyttja extern data från källor som SCB och andra relevanta databaser kan vi få insikter i begagnatbilsmarknadens trender och konsumentbeteenden. Genom att analysera data om försäljningsvolymer, prisutveckling, fordonsmärken och modeller kan vi identifiera vilka bilar som är mest efterfrågade och vilka som erbjuder det bästa värdet för köparen.

Denna data är avgörande för utvecklingen av en begagnad bilplattform som kan möta konsumenternas behov och förväntningar. Genom att integrera extern data i plattformens algoritmer och användargränssnitt kan vi skapa en användarupplevelse som är både intuitiv och informativ, vilket hjälper köpare att fatta välgrundade beslut om sina bilköp.



Bilagor 1

```
# Steg 1: Ladda packages och data

install.packages(c("readxl", "leaps", "broom", "caret", "MASS", "ggplot2"))

library(readxl)

library(randomForest)

library(caret)


# file path

file <- "C:/Users/46736/UC/R-programmering/Samlade_bil_data (1).xlsx"



# Read the Excel file

data_bil <- read_excel(file)
```

```
# Steg 2: Splitar datan till training och testing sets
set.seed(123)
index <- createDataPartition(data_bil$Price, p = 0.3, list = FALSE)
training_data <- data_bil[index, ]
testing_data <- data_bil[-index, ]

# Steg 3: Träna  Random Forest regression model
rf_model <- randomForest(Price ~ ., data = training_data)

# Step 4: Utvärdera
predictions <- predict(rf_model, newdata = testing_data)
mae <- mean(abs(predictions - testing_data$Price))
mse <- mean((predictions - testing_data$Price)^2)
rsquared <- 1 - (sum((testing_data$Price - predictions)^2) / sum((testing_data$Price - mean(testing_data$Price))^2))

# Steg 5: Cross-validation
num_folds <- 10
ctrl <- trainControl(method = "cv", number = num_folds)
model <- train(Price ~ ., data = training_data, method = "rf", trControl = ctrl)

# Print the results
print(model)

# Scatter plot of actual vs. predicted prices
```

```
plot(testing_data$Price, predictions, xlab = "Actual Price", ylab = "Predicted Price", main =
"Actual vs. Predicted Prices")

abline(0, 1, col = "red")  # Add a line of equality (ideal prediction)
```

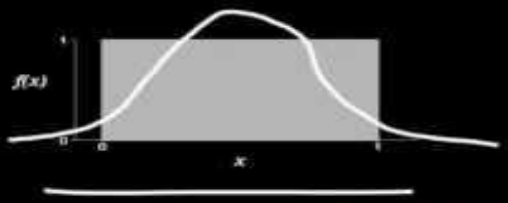
- Källförteckning

https://www.youtube.com/watch?v=a0GjqX5RnxQ&list=PL1DUmTEdeA6LKTMW3wrlT3GiFMCL_r_Sn&ab_channel=%D9%85%D8%AD%D9%85%D8%AF%D8%A7%D9%84%D8%AF%D8%B3%D9%88%D9%82%D9%89

Organize your life and...work with monday.com - the customizable work management platform (youtube.com)

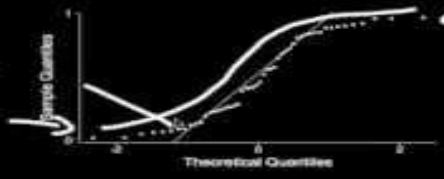
Grundkurs i Excel (youtube.com)

Grundkurs i Excel



Normal QQ plot of random sample of 50 observations:

Smallest values are not as extreme as would be expected



Largest values are not as large (not as extreme) as would be expected

محمد إبراهيم السوقى
المحاضر بقسم نظم المعلومات

جامعة الأمير سطام بن عبد العزيز
Prince Sattam Bin Abdulaziz University
كلية هندسة وعلوم الحاسب
College of Computer Engineering and Science

R

Programming For Beginners