Bacheloroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

Martynas Jocius
Kathrine Sveen

Prisprediksjon for bruktbilsalg: En casestudie med fokus på Nissan Leaf

Bacheloroppgave i Økonomi og administrasjon
Veileder: Denis Becker
April 2024

NTNU
Kunnskap for en bedre verden

Martynas Jocius
Kathrine Sveen

Prisprediksjon for bruktbilsalg: En casestudie med fokus på Nissan Leaf

Bacheloroppgave i Økonomi og administrasjon
Veileder: Denis Becker
April 2024

Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

NTNU
Kunnskap for en bedre verden

Forord

Denne bacheloroppgaven markerer avslutningen på en krevende og spennende studietid ved NTNU Handelshøyskolen. Arbeidet i denne oppgaven viser til vår utvikling gjennom tre studieår, og er et produkt av våre tilegnede kunnskaper innenfor fagfeltene økonomi og informatikk.

Oppgaven tar for seg en problemstilling som vi anser som svært nyttig for det elektriske bruktbilmarkedet fremover. Innholdet i denne oppgaven står for forfatternes regning.

Vi ønsker å benytte anledningen til å takke vår veileder, Denis Becker, for god veiledning og et godt samarbeid i løpet av det siste året. Takk for god konstruktiv kritikk og veiledning, og takk for at du har motivert oss til å utarbeide en best mulig bacheloroppgave. Vi ønsker også å takke Terje Dahlgren og Finn.no for å dele data fra bruktbilsalg med oss. Denne oppgaven hadde ikke vært mulig å gjennomføre uten tilgang til denne eksklusive dataen.

Sammendrag

Denne bacheloroppgaven undersøker ulike statistiske modeller for å kunne predikere priser på brukte Nissan Leaf biler i Norge. Modellene som har blitt benyttet i analysen er lineær regresjon, decision tree regresjon og nevrale nettverk. Resultatene for de ulike modellene blir sammenlignet for å se hvilken som gir de beste predikerte salgsverdiene. For å lage modellene har vi brukt et datasett fra Finn.no med priser fra tidligere salgsannonser. Dette datasettet har vi renset og omkodet før bruk i analysene. Videre har vi laget ulike varianter av datasettet der vi har isolert andre generasjon av Nissan Leaf i ett datasettet, og utelukket datapunkter med verdier større enn 1 i «ad count». Dette har gitt ulike utslag i modellene våre der vi fikk gode resultater med for eksempel nevrale nettverk, der den beste modellen fikk en R^2 på 89,94 %.

Etter drøfting og sammenligning er konklusjonen at nevrale nettverk gir de beste prediksjonene. Samtidig har den flere svakheter i form av underestimering, og tilnærmet ingen predikerte salgsverdier over kr 325 130. Vi har også funnet indikasjoner på at justeringene i datasettet med tanke på variabelen «ad_count» har positiv effekt på prediksjonene til modellene.

Abstract

This bachelor's thesis intends to use various statistical models to be able to predict prices for used Nissan Leaf cars in Norway. The models that have been used are linear regression, decision tree regression and neural networks, and we have compared these against each other to see which ones give the best predicted sales values. We have used a dataset from Finn.no to create the models, with sales prices from previous advertisements. We have cleaned and rearranged this dataset for it to be suitable for the analysis. Furthermore, we have created different variants of the data set where we have isolated the second generation of the Nissan Leaf in one data set, and excluded entries with “ad count” values higher than 1 in another. This has had different results in our models where we got good results in, for example, our neural network model with an R^2 score of 89.94%.

After the analysis and comparisons, it is concluded that neural networks have had the best predicted values. At the same time, it has several weaknesses in the form of underestimation and virtually no predicted sales values above 325 130 NOK. Furthermore, the third data set has indicated towards the most precise predictions in the three models we have used.

Innholdsfortegnelse

Forord .....
Sammendrag .....
Abstract .....
Oversikt over tabeller og figurer .....
1. Innledning 1
2. Teori 2
2.1 Utvikling Nissan Leaf 2
3. Data 3
3.1 Beskrivelse av datasettet 3
3.2 Problemstillinger knyttet til datasettet 3
3.3 Forberedelser av dataen og validering 3
3.3.1 Rensing av datasett 4
3.2.2 Omkoding av kategoriske variabler 5
3.3 Definerer datasettene 5
4. Metode 6
4.1 Statistiske begreper 6
4.2 Lineær regresjon 7
4.3 Decision tree regresjon 8
4.4 Kunstig nevrale nettverk 8
4.4.1 Feedforward nettverk 10
4.4.2 Rectified Linear Unit (ReLU) 10
5. Resultater og drøfting 11
5.1 Modell 1 – Lineær regresjon 11
5.1.1 Datasett 1 11
5.1.2 Datasett 2 13
5.1.3 Datasett 3 15
5.2 Modell 2 – Decision Tree regresjon 18
5.2.1 Datasett 1 18
5.2.2 Datasett 2 20
5.2.3 Datasett 3 22
5.3 Modell 3 – Nevrale nettverk 24
5.3.1 Datasett 1 24
5.3.2 Datasett 2 26
5.3.3 Datasett 3 28
5.4 Oppsummerende drøfting 30

5.4.1 Datagrunnlag 30
5.4.2 Modell 31
6. Konklusjon 33
Referanser 35
Vedlegg 37
Vedlegg 1 37
Vedlegg 2 38
Vedlegg 3 39
Vedlegg 4 – Kode 40

Oversikt over tabeller og figurer

Tabell 5.1: Evalueringsstatistikk for lineær regresjon, modellens prestasjon på testsettene 11
Tabell 5.2: Evalueringsstatistikk for decision tree regresjon, modellens prestasjon på testsettene 18
Tabell 5.3: Evalueringsstatistikk for nevrale nettverk, modellens prestasjon på testsettene 24
Tabell 5.4: Evalueringsstatistikk for modellene med datasett 2 30
Tabell 5.5: Evalueringsstatistikk for modellene med datasett 1 31
Tabell 5.6: Evalueringsstatistikk for modellene med datasett 3 31

Figur 4.1: Eksempel på oppbygningen av et enkelt nevralt nettverk med 2 "hidden layers" 8
Figur 5.1: Spredningsplott over avvik ved prediksjon på testsettet med LR1 12
Figur 5.2: Fordeling over de faktiske prisene i testsettet fra datasett 1 12
Figur 5.3: Fordeling over de predikerte prisene med LR1 på testsettet 12
Figur 5.4: Fordeling over residualene for prediksjonene med LR1 13
Figur 5.5: Spredningsplott over avvik ved prediksjon på testsettet med LR2 14
Figur 5.6: Fordeling over de faktiske prisene i testsettet fra datasett 2 14
Figur 5.7: Fordeling over de predikerte prisene med LR2 på testsettet 14
Figur 5.8: Fordeling over residualene for prediksjonene med LR2 15
Figur 5.9: Spredningsplott over avvik ved prediksjon på testsettet med LR3 16
Figur 5.10: Fordeling over de faktiske prisene i testsettet fra datasett 3 16
Figur 5.11: Fordeling over de predikerte prisene med LR3 på testsettet 16
Figur 5.12: Fordeling over residualene for prediksjonene med LR3 16
Figur 5.13: Spredningsplott over avvik ved prediksjon på testsettet med DTR1 19
Figur 5.14: Fordeling over de faktiske prisene i testsettet fra datasett 1 19
Figur 5.15: Fordeling over de predikerte prisene med DTR1 på testsettet 19
Figur 5.16: Fordeling over residualene for prediksjonene med DTR1 20
Figur 5.17: Spredningsplott over avvik ved prediksjon på testsettet med DTR2 21
Figur 5.18: Fordeling over de faktiske prisene i testsettet fra datasett 2 21
Figur 5.19: Fordeling over de predikerte prisene med DTR2 på testsettet 21
Figur 5.20: Fordeling over residualene for prediksjonene med DTR2 22
Figur 5.21: Spredningsplott over avvik ved prediksjon på testsettet med DTR3 22
Figur 5.22: Fordeling over de faktiske prisene i testsettet fra datasett 3 23
Figur 5.23: Fordeling over de predikerte prisene med DTR3 på testsettet 23
Figur 5.24: Fordeling over residualene for prediksjonene med DTR3 23
Figur 5.25: Spredningsplott over avvik ved prediksjon på testsettet med NN1 25
Figur 5.26: Fordeling over de faktiske prisene i testsettet fra datasett 1 25
Figur 5.27: Fordeling over de predikerte prisene med NN1 på testsettet 25
Figur 5.28: Fordeling over residualene for prediksjonene med NN1 26
Figur 5.29: Spredningsplott over avvik ved prediksjon på testsettet med NN2 27
Figur 5.30: Fordeling over de faktiske prisene i testsettet fra datasett 2 27
Figur 5.31: Fordeling over de predikerte prisene med NN2 på testsettet 27
Figur 5.32: Fordeling over residualene for prediksjonene med NN2 28
Figur 5.33: Spredningsplott over avvik ved prediksjon på testsettet med NN3 28
Figur 5.34: Fordeling over de faktiske prisene i testsettet fra datasett 3 29
Figur 5.35: Fordeling over de predikerte prisene med NN3 på testsettet 29
Figur 5.36: Fordeling over residualene for prediksjonene med NN3 29

1. Innledning

I Norge finner vi en stor eierandel av personbiler med 2 907 164 registrerte personbiler i 2022 (SSB, 2022). Samtidig ser vi også en sterk vekst i registrerte elektriske biler der antall registrerte elektriske personbiler var 599 169 i 2022. Dette tilsvarer en vekst på 331,1% de siste 5 årene (SSB, 2022). Med en slik vekst øker også viktigheten av å ha så riktige prisvurderinger som mulig; både for privatpersoner og for næringslivet. Dette er motivasjonen for denne oppgavens formål; vi ønsker å utforske ulike prediksjonsmodeller som vil kunne tilrettelegge for riktige salgspriser for både privatpersoner og ulike bilsalgsforretninger. Dette vil spesielt være viktig for privatpersoner fremover, da det ikke lengre vil være lov å selge biler «som den er» (Ege, 2024). Ved gode verdivurderinger vil avvikene kunne minimeres, og dermed bidra til mindre risiko vedrørende salg både på privat og profesjonelt nivå.

I denne konteksten har vi valgt ut tre modelltyper som vi ønsker å utforske: lineær regresjon, decision tree regresjon og nevrale nettverk. Disse modellene har hver sine fordeler og ulemper, som vil bli drøftet i løpet av analysene i denne besvarelsen.

For å avgrense omfanget av analysene har vi valgt ut én bestemt bilmodell; Nissan Leaf, og problemstillingen vår lyder derfor som følger:

Hvilken modell og hvilket datagrunnlag gir den beste prediksjonsmodellen til bruk for prising av Nissan Leaf ved bruktbilsalg?

For å kunne predikere priser må vi ta utgangspunkt i tidligere salgspriser. Disse har vi fått tilgang til gjennom Finn.no. Dette datasettet behøvde en del bearbeiding for å kunne brukes i analysearbeider, og har på grunn av dette flere begrensninger. Dette blir utdypet i kapittelet om data.

Denne oppgaven vil ha følgende struktur videre: Først vil vi gjennomgå relevant teori angående bruktbilsalg i Norge. Videre forklarer vi datasettet som har blitt benyttet, problemstillinger ved dataen og justeringene vi har foretatt, før vi så gjennomgår metodene som blir benyttet i analysene. Til slutt ser vi på resultatene fra analysen, og drøfter disse med hensyn på problemstillingen.

2. Teori

SSB viser til at nordmenn, og europeere, bruker bil svært mye relativt stett (SSB, 2017). Ved kjøp av biler er det flere faktorer som er viktig for kunden, altså privatpersonen. Dette inkluderer sikkerhet, vedlikehold og funksjoner, men ikke minst prisen. Viktigheten av pris har blitt desto viktigere i dagens økonomiske situasjon der privatpersoners økonomi er under press av faktorer som høy inflasjon, høye rente kostnader og mer (Sørensen, 2023).

2.1 Utvikling Nissan Leaf

Når vi skal utføre kvantitative analyser så er det nyttig å ha datasett som inkluderer mange datapunkter. Derfor har vi sett på den elektriske bilmodellen Nissan Leaf som har solgt opp mot 80 000 bilmodeller siden lansering i Norge i 2011 (Abrahamsen, 2024). Dette gir oss mange salgspriser spredt utover en lang tidshorisont, noe som er fordelaktig for analysearbeid. Det er samtidig ikke akkurat samme bilmodell som har blitt solgt i 13 år, og derfor vil vi se på utviklingen til Nissan Leaf siden lanseringen i 2011.

Nissan Leaf har produsert ulike bilmodeller siden 2010 med ulike oppgraderinger til både utseendet og ytelse. Under følger en oversikt over de største oppgraderingene som har blitt gjort i forbindelse med lanseringen av nye modeller (Nissan Motor Corporation, 2020):

- 2010: Første modell ble masseprodusert.
- 2012: Oppgradering for batteri, som økte rekkevidden fra 200 til 228 km.
- 2015: Enda større batteri med rekkevidde på 280 km.
- 2017 / 2018: Ny generasjon Nissan Leaf med ny rekkevidde på 400 km.
- 2019: Ny toppmodell, Leaf E+ med 452 km i rekkevidde
- 2020-2022: Oppgraderinger med blant annet raskere lading, og ny Leaf Plus Trim med et større 62 kwh batteri som standard.

Dette kan kategoriseres i to ulike generasjoner, altså første generasjon Nissan Leaf (2010-2017), og andre generasjon Nissan Leaf (2018- i dag).

3. Data

3.1 Beskrivelse av datasettet

Analysene i denne oppgaven er basert på et datasett med informasjon om salg av Nissan Leaf på salgsplattformen Finn.no. Datasettet inkluderer alle annonser for salg av Nissan Leaf fra 2014 til 2023. En oversikt over variablene i datasettet finnes i vedlegg 1.

3.2 Problemstillinger knyttet til datasettet

Før videre beskrivelse av prosessen med å klargjøre datasettet for analysene, er det viktig å presisere hva «alle annonser» betyr i denne sammenhengen. På Finn.no kan man opprette salgsannonser, og dersom produktet, i denne sammenhengen bilen, ikke blir solgt etter en viss tid, justerer man ofte prisen gradvis nedover. Denne handlingen registreres og blir lagret i datasettet under variabelen «ad_count». For hver gang prisen blir justert øker denne variabelen med 1. Samtidig blir den nye prisen på bilen lagt til den gamle prisen og registrert i variabelen «object_price». Dette gjør at variabelen for salgsprisen, der «ad_count» variabelen er større enn 1, er en akkumulert verdi av alle prisene som har vært registrert i den aktuelle salgsannonsen. Dette påvirker påliteligheten til salgsprisene som er oppgitt i datasettet, og er en problemstilling som må håndteres før datasettet blir brukt i analysen. Dette påvirker også variabelen som beskriver antall kilometer bilen har kjørt på samme måte.

En annen problemstilling med måten dataen blir lagret på, er at bilforhandlere som selger via Finn.no ofte gjenbruker annonsene sine. Det vil si at de i stedet for å slette en annonse etter et salg bare går inn på annonsen og oppdaterer dataen med tall for en annen bil de vil selge. Dette blir fanget opp av variabelen «Unique Cars», som øker med 1 for hver unik bil som har blitt registrert i den annonsen. Problemet med dette er at det igjen skaper en oppsamlingseffekt for variablene «object_price» og «mileage». Ettersom det her ikke bare gjelder en oppsamling av data for én unik bil, men for flere ulike biler, skaper dette også problemer for som må håndteres før analysearbeidet.

3.3 Forberedelser av dataen og validering

Det originale datasettet (før rensing) inneholdt 41 104 datapunkter, og inkluderte 13 variabler med ulike formateringer. Formatet på det originale datasettet var ikke enkelt kompatibelt med analyser i Python, så det krevde en del omstrukturering i Excel før datasettet kunne benyttes til selve analysen.

I tillegg til dette måtte flere av variablene omkodes, og en del av datapunktene ble fjernet på grunn av de problemstillingene som ble nevnt i kapittel 3.2.

3.3.1 Rensing av datasett

Datasettet inneholdt noen datapunkter som hadde åpenbare feil og mangler. Dette omfattet datapunkter som manglet registreringsnummer på bilen, og datapunkt med nullverdier på variablene «Unique Cars», «object_price» og «alive_days_estimated». Til sammen utgjorde omtrent 2000 datapunkt. Videre ble en håndfull datapunkter der vi oppdaget manglende samsvar mellom pris og f.eks. «mileage» eller alder på bilen («salgsår» - «year_model») slettet. Disse ble oppdaget gjennom en systematisk gjennomgang av topp- og bunnverdiene for hver variabel, da vi så for oss at det kunne være utliggere der.

Som nevnt var det utfordringer knyttet til datapunkt med akkumulerte verdier for flere ulike biler på grunn av «annonse-praksisen» blant selgere og Finn.no sin metode for dataregistrering. Omfanget av slik «problem-data» var omtrent 6000 datapunkt, og samtlige av disse datapunktene ble derfor slettet, slik at datasettet kun inneholdt annonser med én unik bil per datapunkt.

Etter denne utrensingen besto datasettet fortsatt av over 5000 datapunkt med «ad_count» variabelen er større enn 1. For disse datapunktene ble «object_price» dekomponert ved å dividere de akkumulerte verdiene på verdien av «ad_count», for å simulere en slags gjennomsnittsverdi. Her er det dog viktig å påpeke at alle disse gjennomsnittverdiene for salgsprisen da vil være kunstig høyere enn den reelle salgsprisen ettersom det er naturlig å anta at endelig salgspris var den laveste av de beløpene som har blitt akkumulert.

Variabelen «mileage» ble også konvertert på samme måte, men vi antar her at det ikke er et like stort problem med tanke på forskyvning av verdiene ettersom de fleste bilene ikke nødvendigvis blir kjørt mellom hver gang selgeren justerer salgsprisen. Derfor vil disse gjennomsnittsverdiene være tilnærmet lik de reelle verdiene i de aller fleste tilfellene.

Til slutt ble en del variabler uten variasjon i verdiene (homogenitet), og som derfor ikke ville ha tilført noen verdi i analysearbeidet, slettet. Til sammen utgjorde dette de 5 variabler.

Det ferdige datasettet ble til slutt bestående av 31 203 datapunkt og 8 variabler (county, sales_channel, year_model, sales_year, ad_count, alive_days, object_price, mileage). Hvorav object_price og mileage er de justerte variablene.

3.2.2 Omkoding av kategoriske variabler

Variablene «county» og «sales_channel» måtte omkodes til nummererte kategorier før de datasettet kunne bli benyttet i analysene. En oversikt over de nye verdiene finnes i vedlegg 2.

3.3 Definerer datasettene

I analysen vil de tre utvalgte modelltypene bli trent og testet på tre ulike datasett. Datasett 1 er det ferdig rensede datasettet. Datasett 2 inneholder kun biler fra 2. generasjon av Nissan Leaf; altså kun 2017- til 2023-modeller. Datasett 3 inneholder kun biler med «ad_count» lik 1, som vil si at disse datapunktene ikke blir berørt av problemstillingen rundt akkumulerte verdier og håndteringen av disse. Årsaken til denne inndelingen er at vi ønsker å se om modellene blir mer presise ved å begrense variasjon eller feilkilder i datasettet.

4. Metode

4.1 Statistiske begreper

Residual

Differansen mellom faktisk verdi og predikert verdi; $y_i - \hat{y}_i$ (Hastie et al., 2013, s. 62).

Determinasjonskoeffisienten, R^2

Definisjon:

$$R^2 = \frac{TSS-RSS}{TSS} = 1 - \frac{RSS}{TSS},$$

hvor $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ og $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$

(RSS: residual sum of squares, TSS: total sum of squares)

y_i: $faktisk\ verdi\ for\ observasjon\ i, \hat{y}_i$: $predikert\ verdi\ for\ observasjon\ i$

$\bar{y}$: $gjennomsnittlig\ verdi\ for\ den\ avhengige\ variabelen$

(Hastie et al., 2013, s. 69-70)

Determinasjonskoeffisienten uttrykker hvor stor andel av variasjonen i den avhengige variabelen modellen klarer å fange opp. R^2 tar verdier fra 0 til 1, hvor 0 tilsvarer at modellen ikke fanger opp noe av variansen, mens 1 vil si at modellen fanger opp all variasjon i den avhengige variabelen. Årsaker til lav R^2 kan være at de uavhengige variablene i modellen ikke korrelerer med den avhengige variabelen eller at den valgte modellen ikke er en god tilpasning, eller en kombinasjon av begge (Hastie et al., 2013, s. 80).

Gjennomsnittlig kvadratisk feil (mean squared error), MSE

Definisjon:

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

Jo nærmere de faktiske verdiene prediksjonene er, desto lavere blir MSE. Det er denne verdien som blir brukt ved optimering av modeller; koeffisientene blir tilpasset slik at MSE blir minimert. Videre kan man beregne MSE på test-data for å se hvor godt den tilpassede modellen presterer på ukjent data, og bruke denne verdien for å sammenligne ulike modeller med datagrunnlag med samme fordelingsegenskaper (Hastie et al., 2013, s. 30).

RMSE (root mean squared error)

Definisjon:

$$RMSE = \sqrt{MSE}$$

RMSE kan beskrives som gjennomsnittlig størrelse på residualene. Fordelen med å beregne roten av MSE er at man får verdier i samme enhet og størrelsesorden som datapunktene, og det er derfor lettere å tolke hva denne feilen betyr i praksis (Johnson og Kuhn, 2013, s. 95).

Gjennomsnittlig absolutt feil (mean absolute error), MAE

Definisjon:

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

Et måltall med samme formål som MSE; å gi uttrykk for hvor godt modellen treffer den faktiske dataen. Jo lavere MAE, desto bedre er modellen tilpasset treningsdataen (Hyndman og Athanasoupolus, 2021, kap. 5.8). MAE gir, i likhet med RMSE, verdier med samme enhet som datapunktene.

Gjennomsnittlig absolutt prosentmessig feil (mean absoulte percentage error), MAPE

Definisjon:

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

MAPE uttrykker gjennomsnittlig absolutt avvik i prosent. Ved å relatere størrelsen på avvikene til den faktiske størrelsen på de originale datapunktene er det enklere å si noe om hvorvidt de beregnede avvikene er betydelige i størrelsesorden. Lav MAPE vil si lav andel avvik (Hyndman og Athanasoupolus, 2021, kap. 5.8). Er MAPE eksempelvis 0,15, betyr det at de predikerte verdiene i gjennomsnitt avviker fra de faktiske verdiene med 15%.

4.2 Lineær regresjon

I lineær regresjon forsøker man å finne den beste tilpasningen av en rett linje til datapunktene, slik at feilen mellom de observerte verdiene og de predikerte verdiene er minst mulig. Dette gjøres ved å estimere parameterne til linjen, som vanligvis representeres som stigningstallet og konstantleddet. I en enkel lineær regresjon med én uavhengig variabel kan modellen uttrykkes som (Montgomery, Peck og Vining, 2012) :

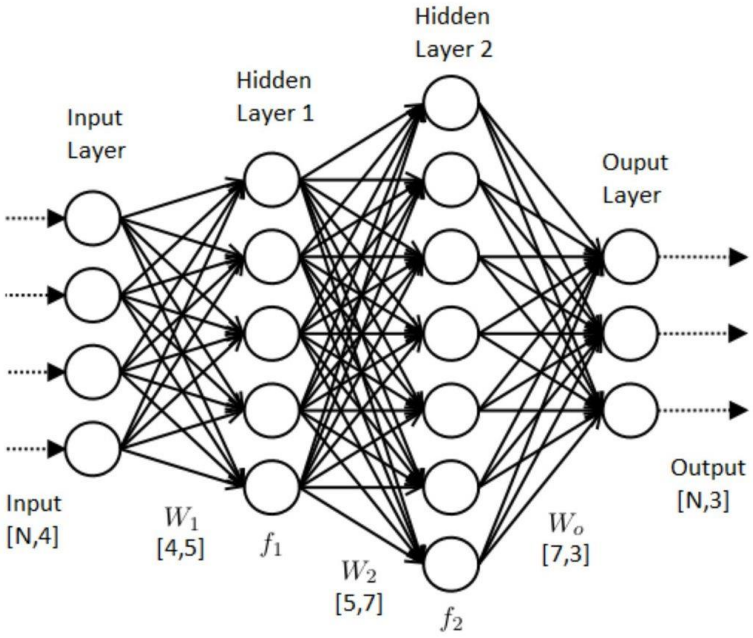
$$Y = \beta 0 + \beta 1 X + \varepsilon$$

Der Y er den avhengige variabelen, X den uavhengige variabelen, $\beta 0$ og $\beta 1$ er parameterne som skal estimeres, og ε er feilleddet (Montgomery, Peck og Vining, 2012).

4.3 Decision tree regresjon

Decision tree regression er en algoritme som bistår med å forutsi kontinuerlige verdier basert på datasett. Dette brukes til å lage et beslutningstre ved å dele dataene inn i mindre grupper, basert på egenskapene til variablene. Hvert tre består av noder som representerer beslutninger, og blader som representerer prediksjonene som kommer til slutt (Breiman, 1984). Treets oppbygning baserer seg på at algoritmen deler datasettet ved å velge de beste splitt funksjonene for å minimere feil. Denne delingen skjer ved å velge den egenskapen som skiller dataen best i hver sin delgruppe (Quinlan, 1986). Når dette treet er bygget, kan vi deretter bruke det til å gjøre prediksjoner. De predikerte verdiene blir gjennomsnittet eller medianen av de faktiske verdiene i bladet (Breiman, 1984). Teknikken ovenfor kan videre tilpasses ved å justere parametere som maksimal dybde, antall datapunkter som kreves for å lage en splitt og mer. Å finne riktig balanse mellom kompleksiteten til modellen og evnen til å generalisere til nye data er viktig (Hastie, Tibshirani og Friedman, 2009).

4.4 Kunstig nevrale nettverk



Figur 4.1: Eksempel på oppbygningen av et enkelt nevralt nettverk med 2 "hidden layers" (Ahire, 2018).

Kunstig nevrale nettverk er en form for dyp maskinlæring, og er oppkalt etter kroppens eget nervesystem.

Nevrale nettverk består av lag med noder. Nodene i ett lag er koblet sammen med nodene i neste lag (se figur 4.1). I noen nettverk er samtlige av nodene i ett lag koblet sammen med hver av nodene i neste lag. Disse beskrives som fullt tilkoblede nettverk. I andre nettverk er det bare noen noder i hvert lag som er koblet sammen med noder i neste lag (Sanderson og Pullen, 2024a). I denne oppgaven benyttes full tilkobling, og teorien videre vil derfor ta utgangspunkt i det.

Hver node i nettverket tar inn en input og gir ut en output. Disse blir representert av verdier mellom 0 og 1. Antallet noder i input laget bestemmes av mengden input-informasjon (Sanderson og Pullen, 2024a). For prismodellen i denne oppgaven, vil dette være antallet uavhengige variabler i datasettet. Fordi nodene kun tar verdier mellom 0 og 1, må dataen fra Finn.no skaleres før den kan benyttes som treningsgrunnlag for det nevrale nettverket. For hver bil i datasettet vil nodene i input-laget bli aktivert med de skalerte verdiene for alle variablene tilhørende bilen. Ettersom en prisprediksjonsmodell skal estimere en kontinuerlig verdi, vil output-laget bare bestå av én node. For en klassifiseringsmodell vil derimot output-laget bestå av én node for hver klassifiseringskategori (Sanderson og Pullen, 2024a).

Lagene imellom input- og output-laget kalles «hidden layers». En modell kan ha alt fra 0 til, i teorien, uendelig mange slike lag. Hvert av disse lagene består av et bestemt antall noder (Sanderson og Pullen, 2024a).

Etter input-laget vil verdien for hver node i de neste lagene bli kalkulert ved hjelp av en aktiveringsfunksjon. Inputen til denne funksjonen baseres på verdiene fra alle nodene i det foregående laget, samt vektene for hver av de korresponderende linkene mellom den spesifikke noden og alle nodene i det foregående laget, og et bias (Sanderson og Pullen, 2024a). Med denne informasjonen beregnes den vektede summen slik:

$$w_{1m} \times x_1 + w_{2m} \times x_2 + \cdots + w_{nm} \times x_n = a_m$$

hvor $w_{1m}, w_{2m}, \ldots, w_{nm}$ er vektene, $x_1, x_2, \ldots, x_n$ er verdiene til nodene i det foregående laget, og a_m er den vektede summen. Før den vektede summen sendes inn som input for aktiveringsfunksjonen, legges det til et bias. Dette blir gjort for å gjøre eventuelle justeringer etter at summen er kalkulert for å sørge for riktig input til aktiveringsfunksjoner. Eksempler på slike justeringer kan være at alle summer under en bestemt grense blir satt til en bestemt verdi (Sanderson og Pullen, 2024a).

Vektende mellom hver node representerer viktigheten, eller styrken, av informasjonen i den foregående noden. Disse kan være både positive og negative. Det er nettopp disse vektene som er parameterne i modellen, og deres verdier blir bestemt gjennom nettverkets trening på data (Sanderson og Pullen, 2024a). Denne treningsprosessen foregår ved hjelp av en algoritme som kalles «the gradient descent algorithm», hvor formålet er å minimere treningsdataens MSE. For hver iterasjon med denne algoritmen kalkuleres MSE med de utvalgte vektene. Algoritmen fortsetter med nye vekter fram til den ikke lenger klarer å forbedre MSE etter et visst antall iterasjoner. I denne oppgaven er 3 iterasjoner brukt som grense (Sanderson og Pullen, 2024b).

4.4.1 Feedforward nettverk

Den enkleste formen for nevrale nettverk kalles Feedforward nettverk (Whitfield, 2022). Dette er et fullt tilkoblet nettverk som karakteriseres ved at treningsprosessen foregår sekvensielt fra ett lag til det neste. Vektene i nettverket starter med tilfeldig valgte verdier. På grunn av at startpunktet for disse vektene nettopp er tilfeldig, vil ulike treninger av den samme modellen kunne gi ulike resultater ved trening på det samme datasettet. Dette er en konsekvens av kompleksiteten av «the gradient descent algorithm». Måten dette har blitt håndtert i denne oppgaven er gjennom klassisk prøving og feiling. Det samme gjelder for valg av lag og noder og lag i modellen. Til slutt endte vi opp med en modell med følgende struktur: 7-5-5-3-1, samt aktiveringsfunksjonen «relu» på alle relevante lag.

4.4.2 Rectified Linear Unit (ReLU)

Definisjon (Aggarwal, 2018, s. 13):

$$\phi(v) = \max\{v, 0\}$$

5. Resultater og drøfting

5.1 Modell 1 – Lineær regresjon

	Datasett 1	Datasett 2	Datasett 3
R^2	0,8536	0,7027	0,8547
MSE	800 755 100	1 008 906 000	800 116 600
RMSE	28 298	31 763	28 286
MAE	21 722	24 481	21 687
MAPE	0,1641	0,1173	0,1640

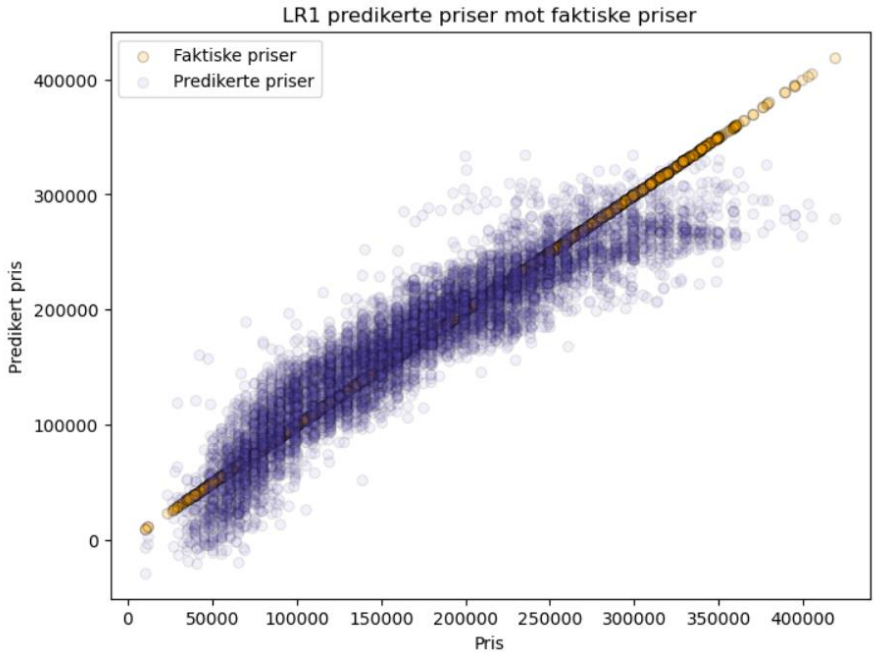
Tabell 5.1: Evalueringsstatistikk for lineær regresjon, modellens prestasjon på testsettene.

Tabell 5.1 gir en oversikt over de statistiske måltallene som vil bli benyttet i analysen for å drøfte hvordan de ulike versjonene av den lineære regresjonsmodellen presterte. Alle måltallene er beregnet ut ifra modellens prestasjon på testsettene for de ulike datasettene. Videre i drøftingen vil den lineære regresjonsmodellen bli omtalt som LR, og de ulike versjonene av modellen LR1, LR2 og LR3, basert på nummereringen på de respektive datasettene modellene er basert på.

5.1.1 Datasett 1

LR1 har en R^2 på 0,8536. Det vil si at modellen forklarer 85,36% av variasjonen i salgsprisen – et tilsynelatende bra resultat, og en god indikasjon på at lineær regresjon kan være en velegnet tilpasning for dataen. MSE for LR1 er på 800 755 100.

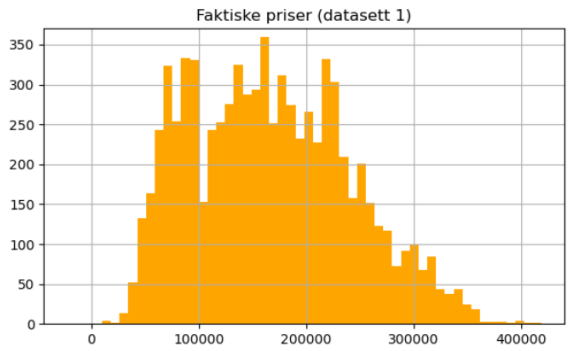
Videre er RMSE lik kr 28 298 og MAE er lik kr 21 722. Det vil si at den gjennomsnittlige differansen mellom prediksjonene og den faktiske dataen er i denne størrelsesordenen, altså omtrent kr 25 000. At RMSE er høyere enn MAE er forventet ettersom avvikene blir kvadrert i beregningen av RMSE, og større avvik dermed blir uttrykt sterkere gjennom dette måltallet enn i MAE, som er basert på absoluttverdier. MAPE på 0,1641 uttrykker at de predikerte verdiene i snitt avviker fra de faktiske med 16,41%.



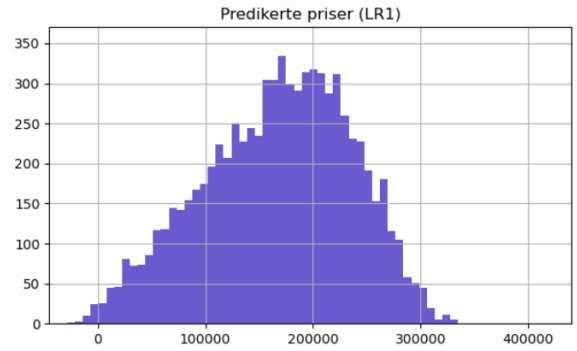
Figur 5.1: Spredningsplott over avvik ved prediksjon på testsettet med LR1.

Figur 5.1 illustrerer avvikene for LR1-modellens prediksjoner på testsettet. Jo lenger unna den oransje diagonalen de lilla prikkene befinner seg, desto større er feilprediksjonen.

Det kommer frem at modellen sliter spesielt med å predikere riktig priser på de dyreste bilene. Modellens høyeste predikerte pris er kr 334 901. Til sammenligning er den faktiske høyeste salgsprisen kr 419 000, og 111 av bilene ble solgt til priser over modellens maksprediksjon. Dette utgjør 1,4% av bilene i testsettet. I tillegg foreslår modellen i flere tilfeller negativ salgspris. Imidlertid ser man også at bilene med priser mellom kr 100 000 og kr 250 000 i stor grad blir overestimert. Ut ifra spredningsplottet kan det dermed virke som at modellen tenderer mot å overestimere prisene på biler som ligger rundt snitt prisen, mens biler med priser mot ytterkantene blir underestimert.

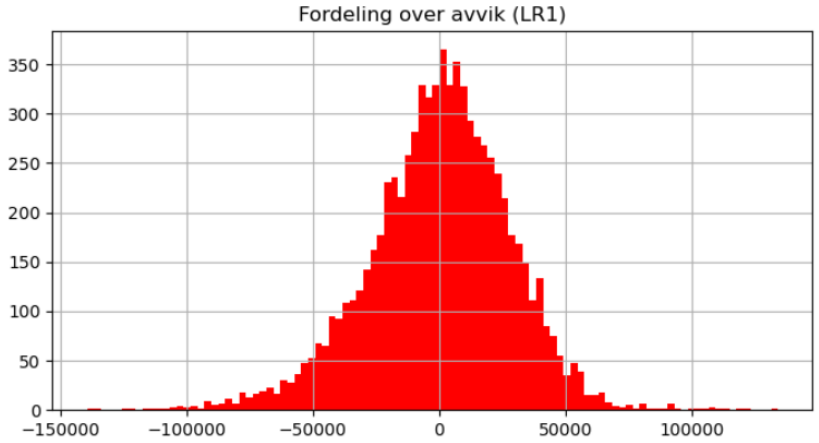


Figur 5.2: Fordeling over de faktiske prisene i testsettet fra datasett 1.



Figur 5.3: Fordeling over de predikerte prisene med LR1 på testsettet.

Sammenligning av fordelingen over de faktiske prisene med fordelingen over de predikerte prisene (figur 5.2 og 5.3), indikerer de samme tendensene som ble observert i spredningsplottet. Gjennomsnittsprisene til de to fordelingene, på henholdsvis kr 167 136 og kr 167 245, gir samsvarende informasjon; til tross for at LR1 både predikerer negative priser, òg kraftig underpriser de dyreste bilene, har de to fordelingene omtrent samme snitt. Dette bekrefter dermed at majoriteten av bilene imellom ytterkantene får overestimerte priser.



Figur 5.4: Fordeling over residualene for prediksjonene med LR1.

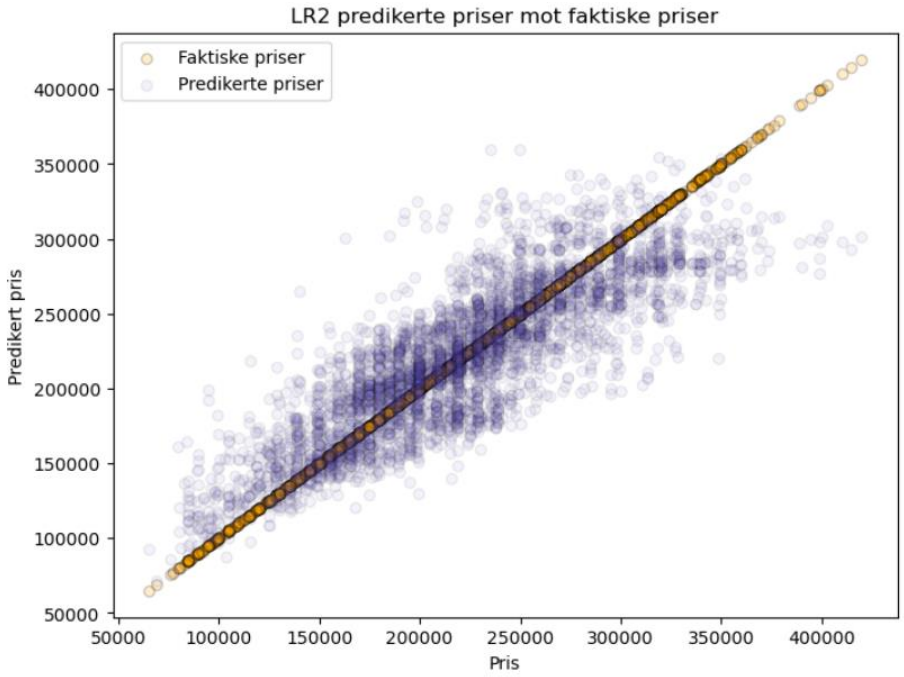
Figur 5.4 illustrerer fordelingen over avvikene for prediksjonene med LR1. Fordelingen har en forventningsverdi på kr 109 (tilsvarende differansen mellom 167 245 og 167 136), noe som også antyder at modellen i snitt vil overestimere prisen på bilene.

LR1 scorer forholdsvis bra på alle statistiske måltall, men ser ikke ut til å være optimal for vårt formål. Årsaken til dette kan både ligge i at salgsprisen trolig ikke er lineært fordelt, eller at datagrunnlaget i form av datasett 1 ikke er gunstig, eller en kombinasjon av begge deler. Dette vil de videre analysene bidra mot å oppklare.

5.1.2 Datasett 2

LR2 har en R^2 på 0,7027. Det vil si at modellen forklarer 70,27% av variasjonen i salgsprisen; en ganske stor nedgang i ytelse fra LR1. Verdiene for RMSE og MAE er henholdsvis kr 31 763 og kr 24 481, og viser til den gjennomsnittlige størrelsen på avvikene mellom predikerte priser og faktiske priser. Disse verdiene peker også mot at LR2 presterer svakere enn LR1.

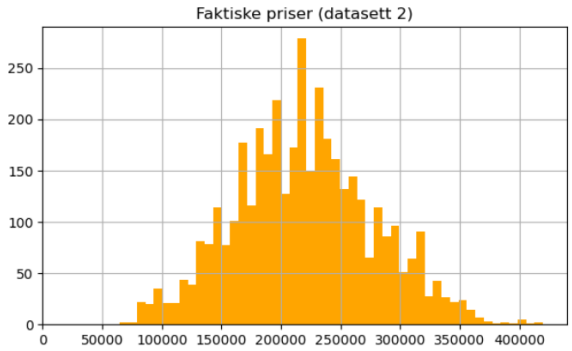
Vi ser likevel en forbedring i MAPE, som har sunket til 0,1173. Dette er overraskende med tanke på resultatene over, og vil derfor bli undersøkt videre i drøftingen.



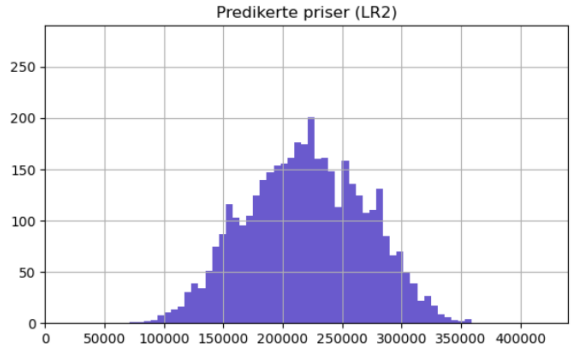
Figur 5.5: Spredningsplott over avvik ved prediksjon på testsettet med LR2.

Figur 5.5 illustrerer avvikene for LR2-modellens prediksjoner på testsettet. Her ser vi med en gang den samme utfordringen som med LR1, at modellen ikke klarer å predikere høye nok priser for de dyreste bilene. Den kommer dog nærmere enn LR1, med høyeste predikerte pris lik kr 359 303.

Der LR1 underestimerte de billigste bilene, gjør LR2 det motsatte, og overestimer de aller fleste bilene i det lavere prissjiktet. For biler med originalpris mellom kr 175 000 og kr 275 000 avviker de predikerte prisene tilsynelatende like mye i positiv som negativ retning, men de største avvikene i dette intervallet går i positiv retning.

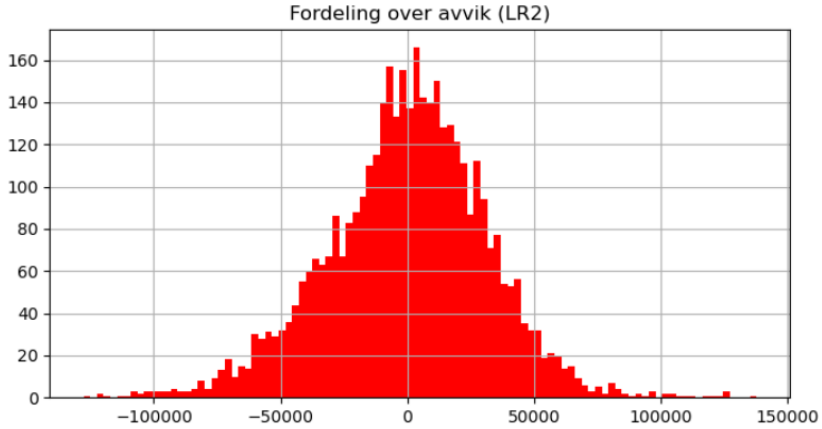


Figur 5.6: Fordeling over de faktiske prisene i testsettet fra datasett 2.



Figur 5.7: Fordeling over de predikerte prisene med LR2 på testsettet.

Fra figurene 5.6 og 5.7 kan en se at LR2 i sum konstruerer en jevnere fordeling over prisene enn de faktisk har i virkeligheten. Den originale fordelingen i datasett 2 er mye spissere, men har samtidig et bredere utfallsrom (mot høyre).



Figur 5.8: Fordeling over residualene for prediksjonene med LR2.

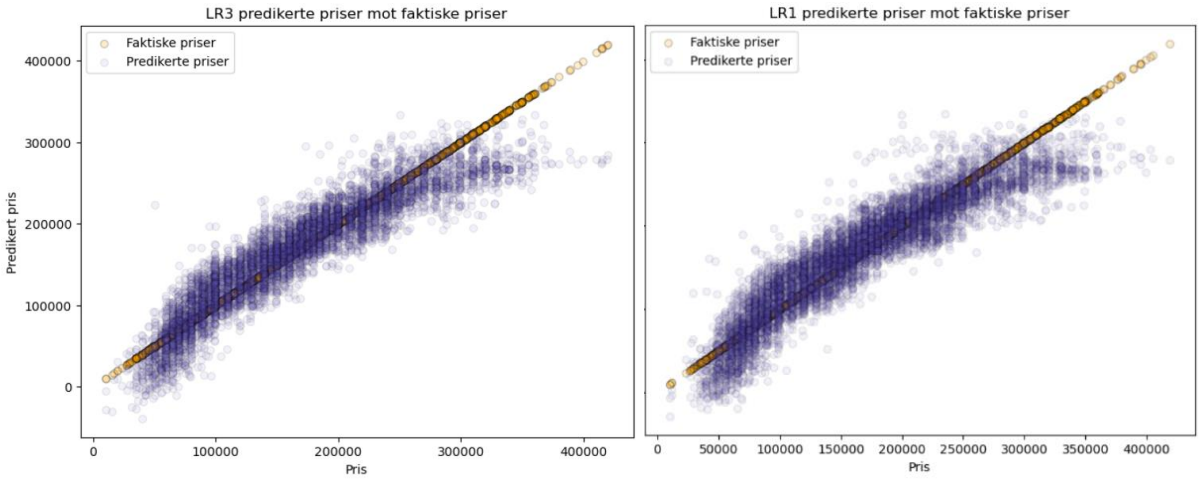
Fordelingen over residualene for prediksjonene med LR2, figur 5.8, har en forventningsverdi på kr -123. Dette gir en indikasjon på at LR2 i snitt underestimerer salgsprisen. Dette stemmer godt overens med observasjonene fra spredningsplottet i figur 5.5 hvor en kunne se at LR2, i likhet med LR1, ikke når opp til de høyeste summene med sine prediksjoner. Det er dog ikke en veldig sterk skjevhet, noe som forklares ved at LR2 overestimerer majoriteten av de resterende (ikke dyreste) bilene.

Oppsummert, bommer LR2 jevnt over litt mer på prediksjonene enn LR1, men til gjengjeld treffer den bedre på ekstremverdiene. Dette kan forklare den overraskende forbedringen i MAPE fra LR1 til LR2, da disse litt større avvikene LR1 gjør på ekstremverdier gir stort utslag ved beregning av prosentmessig snittavvik. Det illustrerer også viktigheten av å bruke flere måltall og å se på fordelingene for modellene når en skal evaluere, da ikke alle evalueringsverdiene klarer å fange opp den samme informasjonen.

Analysen av LR2 indikerer at å bruke data utelukkende fra salg av biler fra 2. generasjon av Nissan Leaf ikke fører til en mer presis modell.

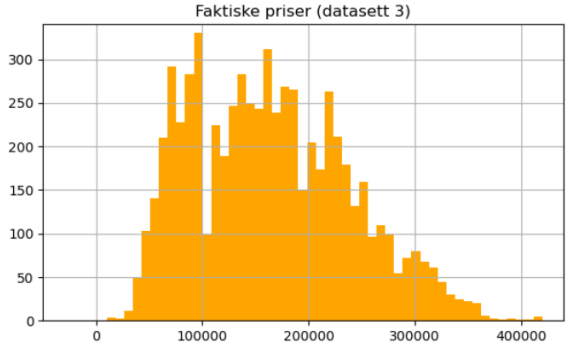
5.1.3 Datasett 3

LR3 presterer generelt veldig likt som LR1 ut ifra måltallene i tabell 5.1.

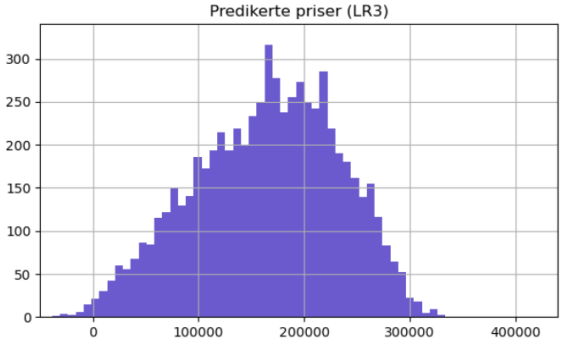


Figur 5.9: Spredningsplott over avvik ved prediksjon på testsettet med LR3 til venstre, spredningsplott over avvik ved prediksjon med LR1.

Spredningsplottene for LR3 og LR1 (figur 5.9) ser ut til å være nesten identiske. LR3 har færre utliggere for prediksjoner mellom ytterpunktene, men avviker dog litt mer for ekstremverdiene. Det er spesielt i det laveste prissjiktet hvor LR3 avviker enda mer i negativ retning enn LR1.

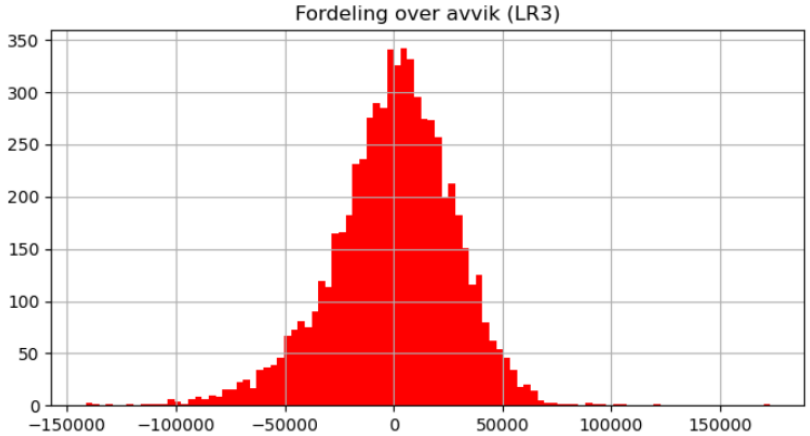


Figur 5.10: Fordeling over de faktiske prisene i testsettet fra datasett 3.



Figur 5.11: Fordeling over de predikerte prisene med LR3 på testsettet.

Figur 5.10 og 5.11 gjenspeiler også de samme observasjonene – det er liten forskjell mellom resultatene for LR1 og LR3.



Figur 5.12: Fordeling over residualene for prediksjonene med LR3.

Fordelingen over residualene for LR3 i figur 5.12 bekrefter dog den forskjellen som ble observert ved sammenligning av spredningsplottene til LR1 og LR3; at LR3 har en endra sterkere forskyving i negativ retning for ekstremverdiene. Der figur 5.4 hadde en forventningsverdi på kr 109, har figur 5.12 en forventningsverdi på kr -252. Det er likevel viktig å påpeke at dette er såpass små verdier i forhold til nivået på salgsprisene at det ikke er grunnlag for å påstå at den ene modellen overvurderer og den andre undervurderer.

Til tross for tilsynelatende gode verdier på de statistiske måltallene, viser nærmere analyser av prediksjonene fra LR1, LR2 og LR3 at lineær regresjon ikke er en veldig god tilpasning for dataen. Det kan dog bemerkes at LR1 og LR3 gir veldig like resultater, noe som indikerer at justeringene gjort fra datasett 1 til datasett 3 ikke har utslagsgivende effekt på prestasjonene.

5.2 Modell 2 – Decision Tree regresjon

	Datasett 1	**Datasett 2**	**Datasett 3**
R^2	0,8299	0,6211	0,8446
MSE	930 518 714	1 285 697 071	857 338 009
RMSE	30 504	35 857	29 280
MAE	21 202	25 916	20 448
MAPE	0,1467	0,1208	0,1385

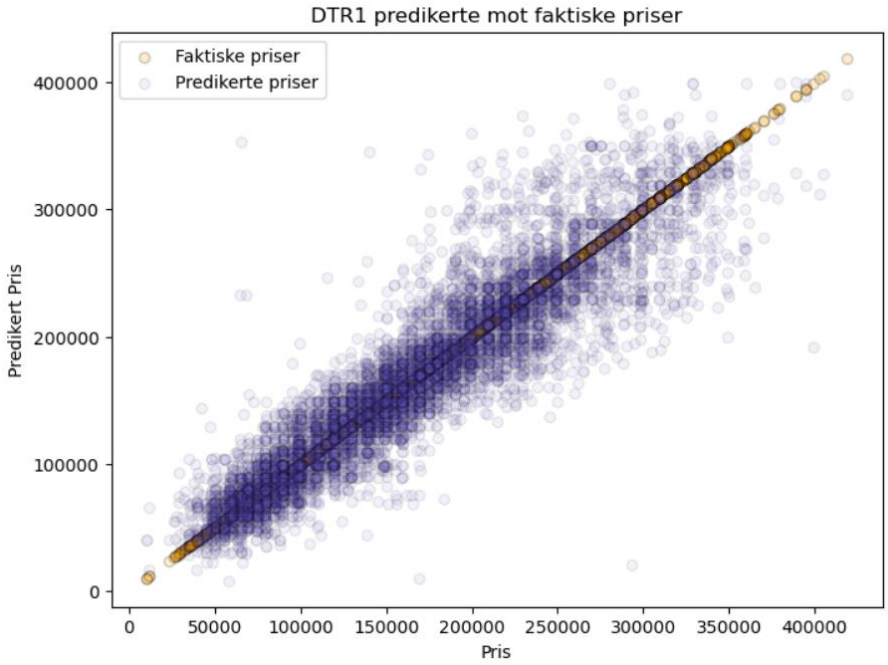
Tabell 5.2: Evalueringsstatistikk for decision tree regresjon, modellens prestasjon på testsettene.

Tabell 5.2 gir en oversikt over måltallene som vil bli benyttet i analysen for å drøfte resultatene fra decision tree regresjon på de ulike datasettene. Resultatene er beregnet ut ifra modellens prestasjon på testsettene for de ulike datasettene. Utover i drøftingen vil denne modellen omtales som DTR, og de ulike versjonene vil bli omtalt som DTR1, DTR2, og DTR3, basert på hvilket datasett har blitt benyttet.

5.2.1 Datasett 1

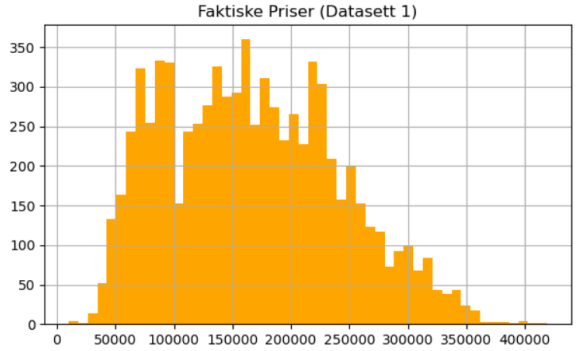
DTR1 har en R^2 på 0,8444. Det vil si at modellen forklarer 84,44% av variasjonen, noe som er et godt resultat, og gir oss en god indikasjon på at decision tree regresjon kan være en gunstig modell for datasettet vårt. MSE er på 930 518 714.

Videre har vi en RMSE på kr 30 504 og en MAE lik kr 21 202. Det forteller oss den gjennomsnittlige differansen mellom prediksjonene og de faktiske salgsprisene i datasettet vårt. Som tidligere antatt i besvarelsen vår så har vi en høyere RMSE enn MAE på grunn av kvadrerte avvik. Til slutt ser vi en MAPE på 0,1467, som gir uttrykk for at de predikerte verdiene avviker i snitt med 14,67 %.

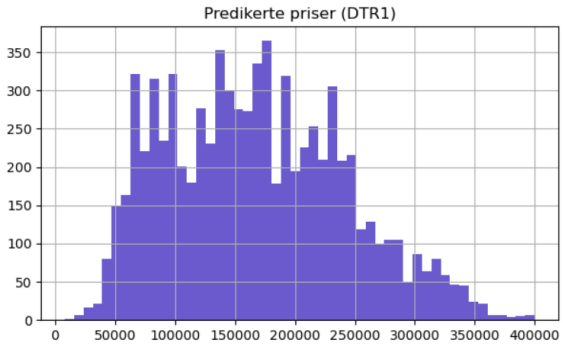


Figur 5.13 2: Spredningsplott over avvik ved prediksjon på testsettet med DTR1.

Figur 5.13 viser avvikene for DTR1-modellens prediksjoner på testsettet fra datasett 1. Spredningsplottet viser store avvik, både i negativ og positiv prisretning. Modellens prediksjoner har en viss vifteform, der den predikerer godt på biler med lavere salgsverdi sammenlignet med biler med høyere salgsverdi.

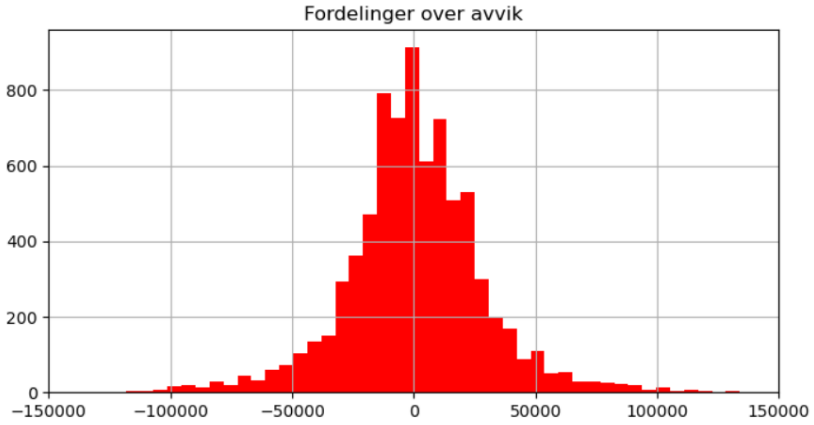


Figur 5.141: Fordeling over de faktiske prisene i testsettet fra datasett 1.



Figur 5.15: Fordeling over de predikerte prisene på testsettet med DTR1.

Ved sammenligning av fordelingen over de faktiske og predikerte prisene (se fig. 5.14 og 5.15) ser vi flere av de samme tendensene som ble observert i spredningsplottet. Disse illustrasjonene viser at modellen predikerer flere salgspriser i det øverste segmentet sammenlignet med de faktiske prisene i datasettet.



Figur 5.16: Fordeling over residualene for prediksjonene med DTR1.

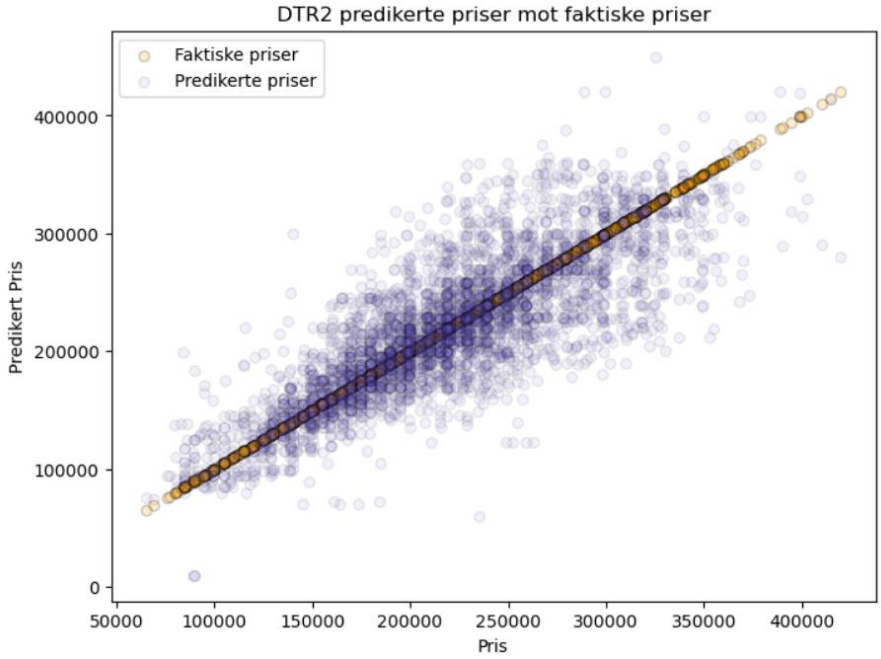
Figur 5.16 illustrerer fordelingen over avvikene for prediksjonene til DTR1. Fordelingen har en forventningsverdi på kr 400 som indikerer at modellen i snitt overestimerer prisene på bilen. Til tross for mange avvik, viser fordelingen at brorparten av avvikene ligger nærme 0. Det er altså mye avvik, men de fleste avvikene er av en lav størrelsesorden.

DTR1 scorer relativt bra på de fleste statistiske måltall, men har også sine svakheter. Svakhetene finner man i de høyere salgsverdiene som varierer svært mye i form av både overestimering og underestimering.

5.2.2 Datasett 2

DTR2 har en R^2 på 0,6211. Det vil si at modellen forklarer 62,11 % av variasjonen i salgsprisen som er en svært stor nedgang i forklaringsevne sammenlignet med DTR1. Verdiene for RMSE og MAE er kr 35 857 og kr 25 916 som viser til det gjennomsnittlige omfanget av avvikene mellom faktiske og predikerte priser. Disse verdiene peker også mot at DTR2 presterer dårligere enn DTR1.

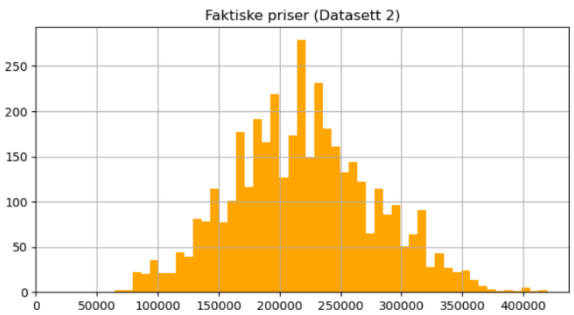
Til tross for dårligere verdier i de fleste statistiske måltall så får vi en lavere MAPE på 12,08 %. Dette vil bli diskutert senrer i drøftingen.



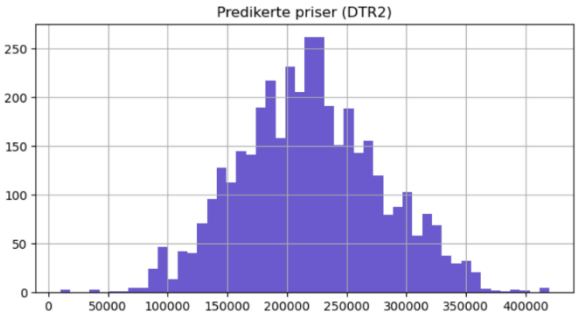
Figur 5.17: Spredningsplott over avvik ved prediksjon på testsettet med DTR2.

Figur 5.17 illustrerer avvikene mellom de predikerte og faktiske prisene i testsettet. Sammenlignet med DTR1 ser vi også en vifteform, men at avvikene vokser raskere med økt salgspris. I tillegg predikerer den færre salgsverdier opp mot det høyeste prissjiktet.

Der DTR1 predikerte relativt nærme faktisk pris for de billigste bilene, inntil kr 250 000 kroner, så gir DTR2 predikerte verdier som tydelig både over- og underestimerer også i dette prissjiktet.



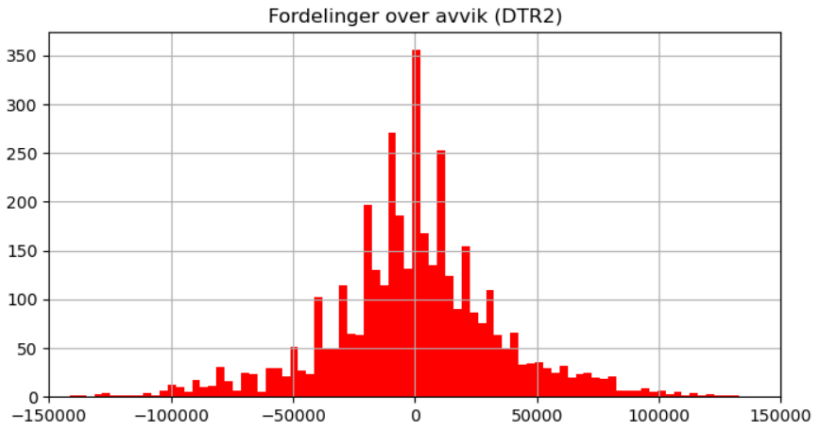
Figur 5.18: Fordeling over de faktiske prisene i testsettet fra datasett 2.



Figur 5.19: Fordeling over de predikerte prisene på testsettet med DTR2.

Figur 5.18 og 5.19 viser fordelingene for de faktiske og de predikerte prisene i datasett 2. Fordelingen over de faktiske prisene i datasett 2 sett i sammenheng med fordelingen i datasett 1, kan være forklaringen på den overraskende forbedringen i MAPE fra DTR1 til DTR2. For datasett 2 er det mindre variasjon i den faktiske dataen, noe som ikke gir like mye rom for store feilprediksjoner. Selv om DTR2 generelt gir upresise prediksjoner, er hvert avvik likevel

ikke veldig store. Vi ser også at prediksjonene til DTR2 enkelt har oppnådd en lignende fordeling som de faktiske dataene til tross for relativt stor spredning i spredningsplottet.



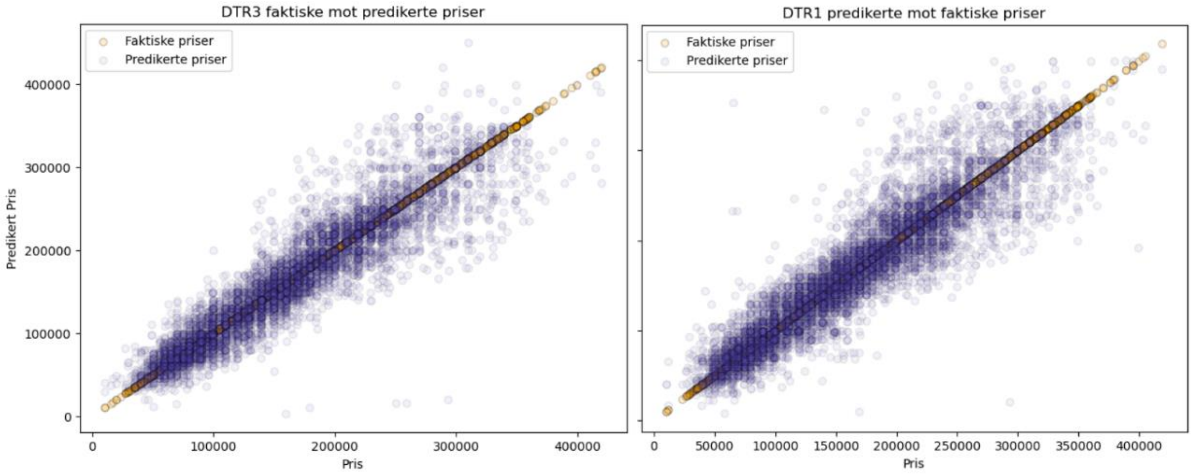
Figur 5.20: Fordeling over residualene for prediksjon med DTR2.

Fordelingen i figur 5.20 har en forventningsverdi på kr -398 som tilsier at DTR2 i snitt underestimerer salgsprisen. Dette samsvarer med observasjonene ovenfor der DTR2 ofte ikke predikerer høyt nok på salgsverdier over kr 350 000. Vi ser færre lave avvik enn det vi finner på DTR1, der avvik på 0 hadde en frekvens på 800, sammenlignet med 350 for DTR2.

DTR2 bommer oppsummert sett mer på prediksjonene sine enn DTR1, der vi ser dårligere verdier i R^2, RMSE og MAE. Dette indikerer at å utelukke første generasjon med Nissan Leaf ikke fører til bedre predikerte verdier.

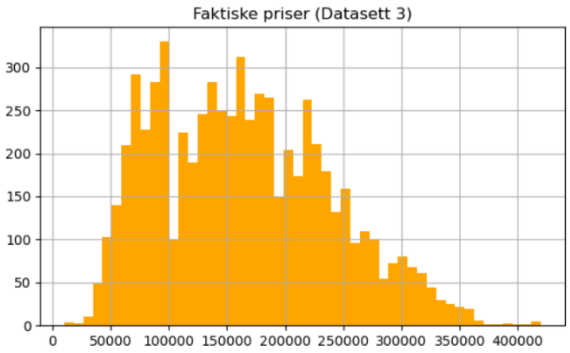
5.2.3 Datasett 3

DTR3 predikerer på svært lik måte som DTR1 ut ifra måltallene i tabell 5.2, med få forbedringer i R^2, RSME, MAE og MAPE.

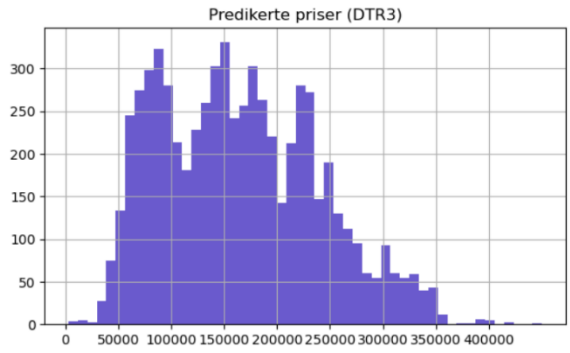


Figur 5.21: Spredningsplott over avvik ved prediksjon på testsettet med DTR3 til venstre, spredningsplott over avvik ved prediksjon med DTR1.

Spredningsplottet i figur 5.21 har samme vifteform som vi fant for DTR1, men med større spredning i intervallet fra kr 200 000 til 250 000.

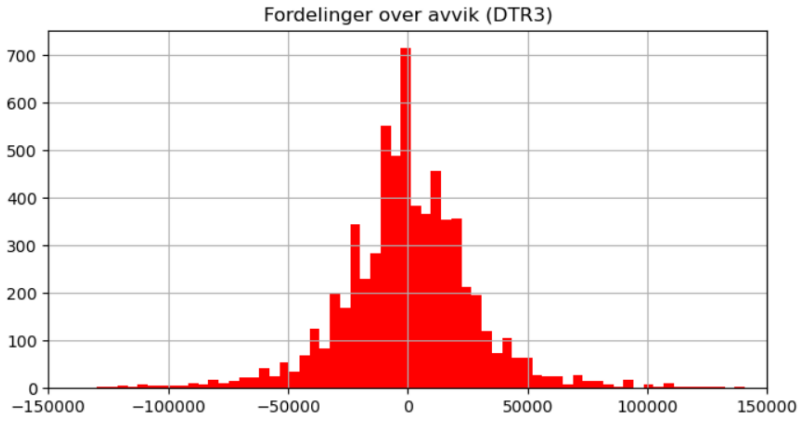


Figur 5.22: Fordeling over de faktiske prisene i testsettet fra datasett 3.



Figur 5.23: Fordeling over de predikerte prisene på testsettet med DTR3.

I figur 5.22 og 5.23 ser vi relativt like observasjoner som for DTR1, som igjen gjenspeiles i de relativt små forskjellene i de statiske måltallene.



Figur 5.24: Fordeling over residualene for prediksjon med DTR3.

Residualfordelingen for DTR3 i figur 5.24 viser en lignende spiss form som DTR1, men med en svakt negativ forventningsverdi på kr -87. Det er samtidig viktig å presisere at det er svært lave verdier i forhold til salgsprisene på Nissan Leaf, og derfor mindre grunnlag for å si at DTR3 undervurderer sammenlignet med DTR1.

De ulike datasettene har gitt forholdsmessig gode statistiske måltall der DTR1 og DTR3 predikerer best. Modellene tyder på en vifteform på spesielt DTR1 og DTR3 som indikerer bedre prediksjoner på biler med lavere salgsverdi enn høy salgsverdi. Forskjellene mellom DTR1 og DTR3 er dog såpass like vi ikke kan fastslå at justeringene som har blitt gjort på datasett 3 i forhold til datasett 1 har hatt en vesentlig effekt på prediksjonene.

5.3 Modell 3 – Nevrale nettverk

	Datasett 1	**Datasett 2**	**Datasett 3**
R^2	0,8977	0,7508	0,8994
MSE	559 638 145	845 656 221	554 216 917
RMSE	23 657	29 080	23 542
MAE	17 428	21 495	17 456
MAPE	0,1189	0,1001	0,1203

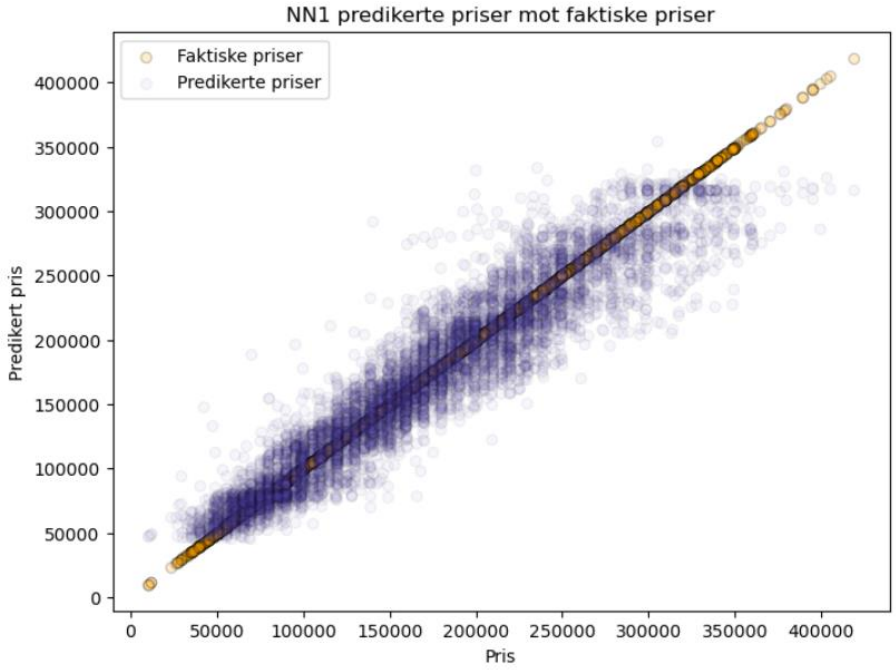
Tabell 5.3: Evalueringsstatistikk for nevrale nettverk, modellens prestasjon på testsettet.

Tabell 5.3 gir en oversikt over de statistiske måltallene som vil bli benyttet i analysen for å drøfte modellene med nevrale nettverk. Alle måltallene er beregnet ut ifra modellens prestasjon på testsettene for de ulike datasettene. Videre i drøftingen vil den nevrale nettverksmodellen bli omtalt som NN, og de ulike versjonene av modellen NN1, NN2 og NN3, basert på nummereringen på de respektive datasettene modellene er basert på.

5.3.1 Datasett 1

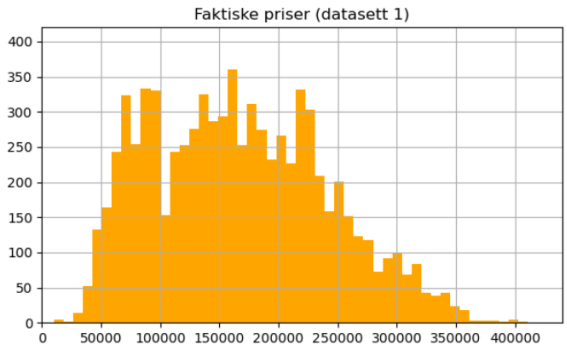
NN1 har en R^2 på 0,8977, som vil si at modellen fanger opp 89,77% av variasjonen i dataen. Dette er en bra score, som sammen med en MSE på 559 638 145 viser at NN1 klarer å tilpasse seg dataen godt.

NN1 har RMSE lik kr 23 657 og MAE lik kr 17 428. Dette indikerer at modellens prediksjoner i snitt avviker fra de faktiske prisene med beløp i denne størrelsesordenen. Forskjellen i RMSE og MAE kan, som nevnt i kapittel 5.1.1 skyldes at datasettet inneholder noen utliggere. Med en snittpris på bilene i datasettet på omtrent kr 167 000 er dette avviket å beregne som akseptabelt. Prosentmessig avviker prediksjonene fra de faktiske verdiene med 11,89%, jf. MAPE.

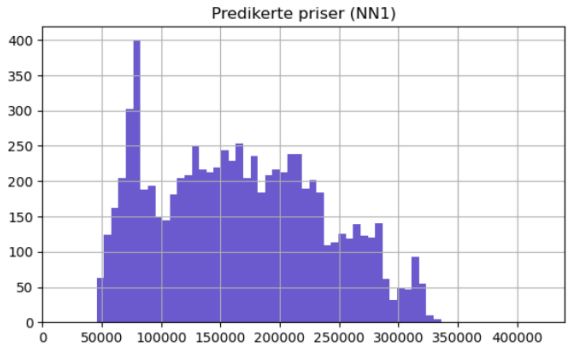


Figur 5.25: Spredningsplott over avvik ved prediksjon på testsettet med NN1.

Spredningsplottet i figur 5.25 viser at prediksjonene har forholdsvis jevnt like store avvik for de fleste prisklasser. Det er dog en tendens til at modellen avviker mer fra faktisk pris etter hvert som salgsprisen stiger. Dette gjør at avvikene har en vifteform, men med en veldig spiss vinkel. Det ser ut til å være en tendens til overprising i det laveste prissjiktet, og en klar tendens til underprising i det høyeste prissjiktet. Modellens høyeste predikerte pris er kr 354 641, mens den faktiske høyeste prisen er, som nevnt tidligere i drøftingen, kr 419 000. Gapet mellom laveste predikerte, og laveste faktiske pris er på kr 35 582 (45 582 - 10 000).



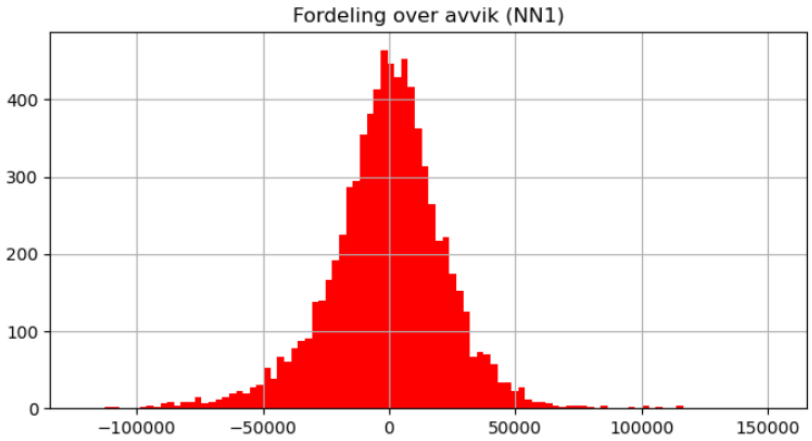
Figur 5.26 3: Fordeling over de faktiske prisene i testsettet fra datasett 1.



Figur 5.27: Fordeling over de predikerte prisene med NN1 på testsettet.

Figurene 5.26 og 5.27 viser hvordan de faktiske og de predikerte salgsprisene fordeler seg. NN1 konstruerer en forholdsvis lik fordeling som den originale, men avvikene som ble fanget opp av spredningsplottet kommer også fram her. Fra figurene kan man også se at de predikerte prisene mellom kr 50 000 og 100 000 er spissere fordelt enn de faktiske prisene i

samme prisintervall. I tillegg er det en større mengde biler med priser mellom kr 275 000 og 325 000 for prediksjonene.



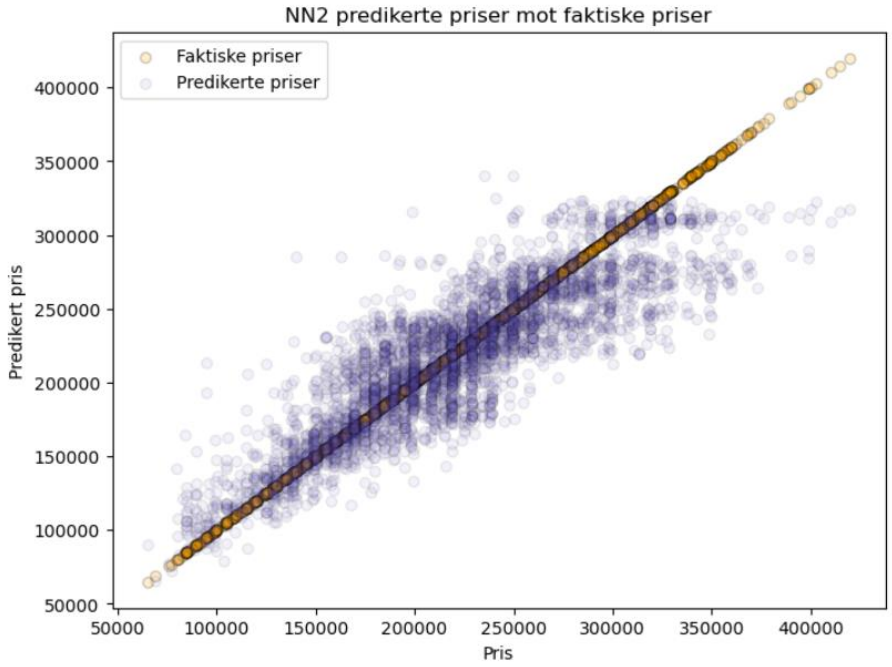
Figur 5.28: Fordeling over residualene for prediksjonene med NN1.

Nærmere analyse av residualene, viser at forventningsverdien til fordelingen over prediksjonsavvikene er kr -817 (figur 5.28). Det peker mot at modellen i snitt har en tendens til å underestimere salgsprisen. Årsaken til dette utfallet ligger nok i modellens utfordring med å nå opp til de høyeste prisene, og at dette gir stort utslag i ulike måltall da størrelsesordenen på disse avvikene i absolutt forstand er veldig stor i forhold til avvikene på laveste prisklassene. Det relative avviket blir ikke fanget opp.

NN1 ser ut til å være en god modell for prediksjon av bruktsalgsprisen til Nissan Leaf, men ikke for bilene i det høyeste og laveste prissjiktet.

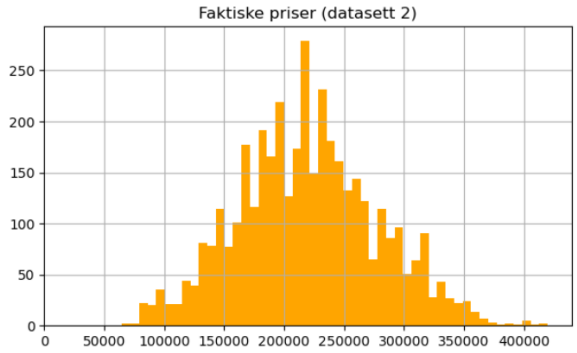
5.3.2 Datasett 2

NN2 har en R^2 på 0,7508; en nedgang på 0,1469 fra NN1. RMSE og MAE indikerer også en nedgang i ytelse fra NN1, med verdier på henholdsvis kr 29 080 og kr 21 495. MAPE er dog noe bedre for NN2, på 0,1001.

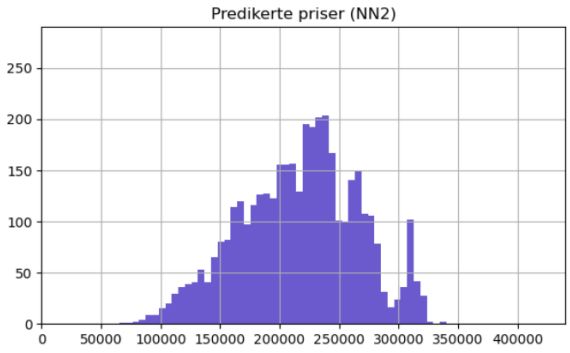


Figur 5.29: Spredningsplott over avvik ved prediksjon på testsettet med NN2.

Figur 5.29 viser spredningsplottet over avvikene mellom prediksjonene til NN2 og de faktiske prisene. Det viser at NN2 jevnt over har større avvik i prediksjonene, men ser til gjengjeld ut til å treffe litt bedre i det laveste prissjiktet enn NN1. Utfordringen med å fange opp de høyeste prisene er også et faktum her, og det ser ut til at modellen «samler» opp de dyreste bilene rundt en bestemt pris – en slags terskel.

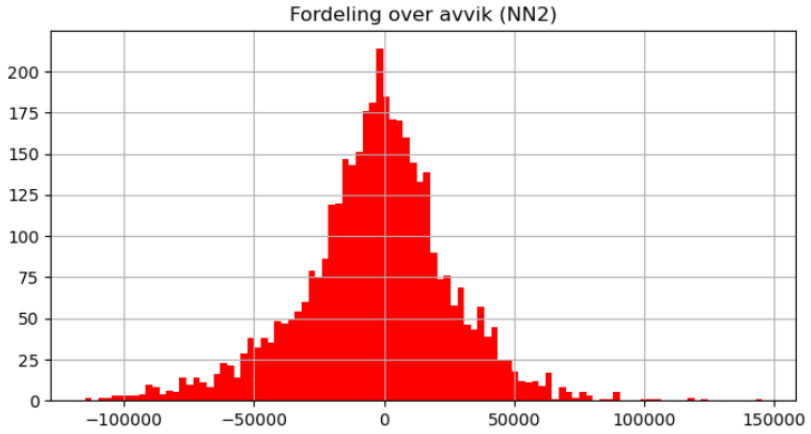


Figur 5.30: Fordeling over de faktiske prisene i testsettet fra datasett 2.



Figur 5.31: Fordeling over de predikerte prisene med NN2 på testsettet.

Figur 5.30 og 5.31 viser at fordelingen over de predikerte prisene har en tykkere hale på venstre side enn fordelingen over de faktiske prisene. Oppsamlingen rundt «terskelen» i det høyeste prissjiktet er også tydelig fanget opp i figur 5.19, hvor den høyre halen kuttes av rundt kr 300 000, og er erstattet av en tydelig spiss med høy frekvens av noen få salgspriser. I midten av fordelingen er de predikerte verdiene fordelt litt jevnere. Dette er nok en naturlig konsekvens av at de faktiske prisene ofte settes til spesifikke tall (f.eks. 99 999 i stedet for 100 000 osv.), mens modellen gir prediksjoner på en kontinuerlig skala.

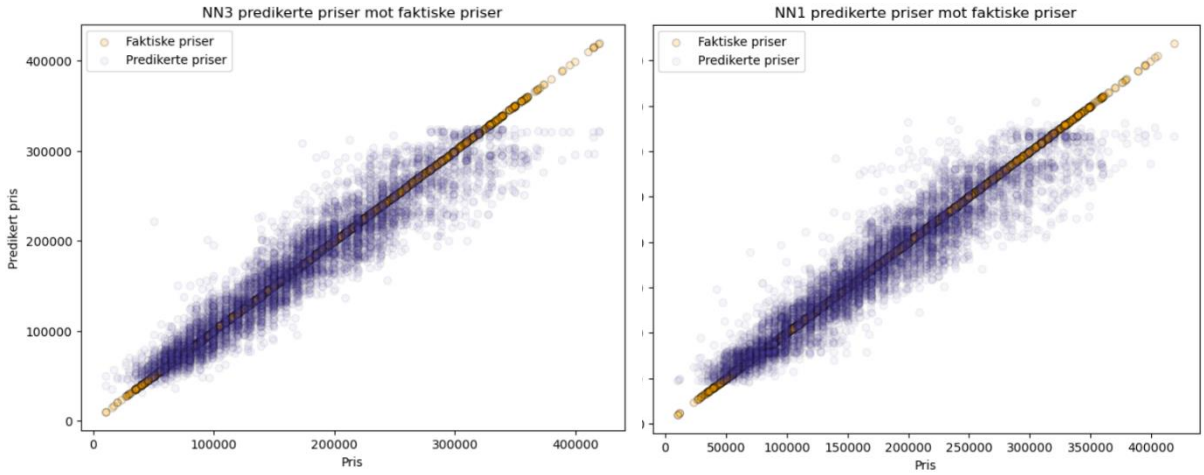


Figur 5.32: Fordeling over residualene ved prediksjoner med NN2.

Fordelingen over residualene for prediksjonene med NN2 i figur 5.32 illustrerer også modellens utfordring i forhold til underprising. Forventningsverdien på kr -3 341, peker klart mot at modellen tenderer til å predikere for lave priser. I den sammenheng, er det imidlertid viktig å tenke over hvorfor modellen får disse resultatene. Modellens problem ligger jo i hovedsak i prediksjonen av bilene i den høyeste prisklassen. Selv om det er veldig store avvik for disse bilene, presterer jo modellen godt i de resterende prisklassene. Denne bemerkningen tar vi med oss videre i den avsluttende drøftingen.

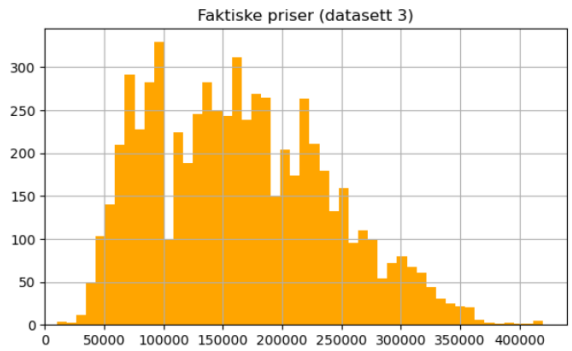
5.3.3 Datasett 3

Resultatene for NN3 er nesten identiske med NN1 (se tabell 5.3). Alle måltall har dog økt, men ikke nok til at den utpeker seg som en tydelig bedre ytende modell.

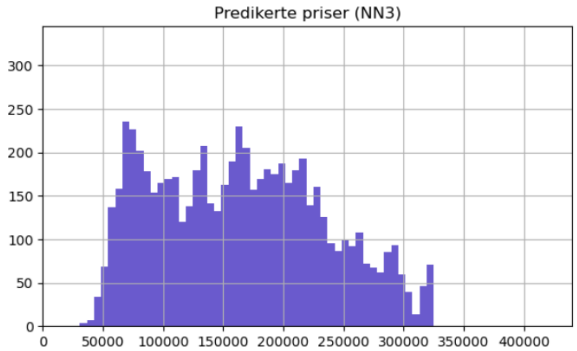


Figur 5.33: Spredningsplott over avvik ved prediksjon på testsettet med NN3 til venstre, spredningsplott over avvik ved prediksjon med NN1.

Spredningsplottet over avvikene for NN3-modellens prediksjoner (figur 5.33) er også veldig likt spredningsplottet for NN1. NN3 ser ut til å ha litt mindre avvik for majoriteten av prediksjonene, men har til gjengjeld en enda lavere makspredikskjon enn NN1, på kr 325 130.

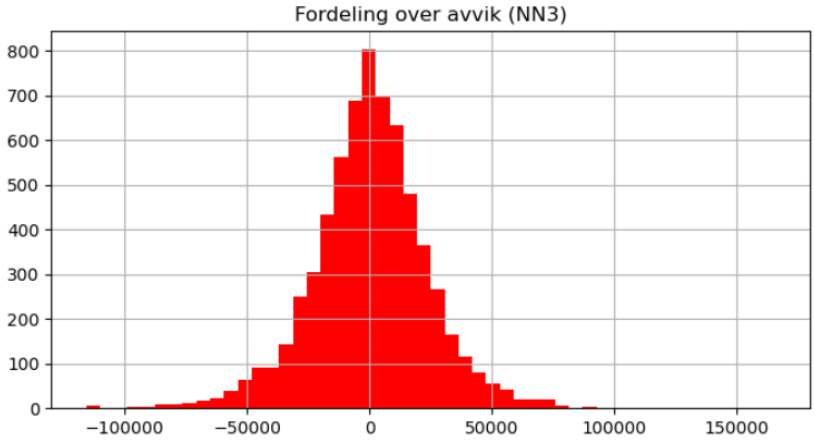


Figur 5.34: Fordeling over de faktiske prisene i testsettet fra datasett 3.



Figur 5.35 4: Fordeling over de predikerte prisene med NN3 på testsettet.

Figur 5.34 og 5.35 viser at NN3 konstruerer en god tilnærming til fordeling over prisene. Den mest utpregede forskjellen i fordelingene er halen på høyre side, hvor man igjen ser at modellen får denne spissen rundt sine høyeste prisprediksjoner. Observasjonene gjort i spredningsplottet ser dermed ut til å stemme veldig bra; NN3 presterer jevnt over veldig godt, men har en stor utfordring i det høyeste prissjiktet.



Figur 5.36: Fordeling over residualene ved prediksjon med NN3.

Fordelingen over residualene for NN3 (se figur 5.36) underbygger også observasjonen om at avvikene fordeler seg jevnt. Den har imidlertid en negativ forventningsverdi på kr -565. Tar man imidlertid høyde for modellens utfordring med det høyeste prissjiktet, kan det antas at residualene for de resterende prediksjonene har en forventningsverdi som er nærmere kr 0, og altså ikke en tendens mot å avvike i negativ retning som figur 5.24 indikerer.

Analysen av NN1, NN2 og NN3 viser at nevrale nettverk evner å tilpasse seg dataen meget godt, og at dette er en god kandidat som prismodell for salg av brukt Nissan Leaf. Den største problemstillingen til alle tre modellene er prediksjoner av priser for bilene i det høyeste prissjiktet. Av de tre modellene er det NN1 og NN3 som presterer best, noe som kan tyde på at datasett 2 ikke er like egnet som datagrunnlag.

5.4 Oppsummerende drøfting

5.4.1 Datagrunnlag

I analysene av de ulike modellene; LR, DTR og NN, får alle modellversjoner med datasett 1 og datasett 3 som treningsgrunnlag, bedre resultater enn de med datasett 2 som grunnlag. Det er derfor rimelig å konstatere at å fjerne de eldste bilene, slik vi har gjort i datasett 2, ikke er fordelaktig. Dette kan ha flere mulige forklaringer. Fra datasett 1 til datasett 2 ble omtrent halvparten av datapunktene fjernet. Dette betyr at modellene som er basert på datasett 2 har fått mye mindre trening, og har derfor mindre mulighet for å få fanget opp variasjonen i dataen.

Det er også viktig å poengtere at den halvdelen av datapunktene som ble tatt ut av datasett 2, er spesifikt de eldste bilene – altså ikke et tilfeldig utvalg datapunkter. Dette gjør at datasett 2 har en ulik fordeling enn de andre to datasettene, og dermed andre egenskaper. Denne forskjellen kan være en del av årsaken til at modellene med datasett 2, til tross for ellers dårlige resultater, likevel har fått best MAPE. Altså at de i snitt bommer minst i sine prediksjoner. Som nevnt i kapittel 5.1.2 har datasett 2 en jevnere fordeling med et smalere utfallsrom enn de andre datasettene, og har derfor mindre rom for større avvik på ekstremverdier.

Når det er sagt, har datasett 2 ikke fått spesielt gode resultater med noen av modellene i denne analysen, ref. tabell 5.4. Hvorvidt dette er på grunn av for få datapunkter, eller at ingen av de utvalgte modellene i denne analysen er god match for datasett 2, har ikke vi grunnlag for å si noe om.

Datasett 2:

	LR	**DTR**	**NN**
R^2	0,7027	0,6211	0,7508
MSE	1 008 906	1 285 697 071	845 656 221
RMSE	31 763	35 857	29 080
MAE	24 481	25 916	21 495
MAPE	0,1173	0,1208	0,1001

Tabell 5.4: Evalueringsstatistikk for modellene med datasett 2.

Det er i utgangspunktet minimale forskjeller mellom prestasjonene til modellene basert på datasett 1 og datasett 3, men modellene med datasett 3 har konsekvent litt bedre resultater enn de med datasett 1 (se tabell 5.5 og 5.6). I de foregående analysene hvor modellene ble evaluert separat, var ikke dette nok grunnlag for å påstå at datasett 3 fører til bedre prestasjoner. Nå

derimot, når vi ser at dette ble utfallet for alle de tre analysene separat, er det større grunn til å anta at utvalget i datasett 3 medfører litt mer presise prediksjoner.

Forskjellen mellom datasett 1 og datasett 3 er at vi har fjernet alle datapunkt som har åpenbare usikkerheter i flere av variablene. I kapittel 3.2 ble ulike utfordringer med dataen gjennomgått, blant annet problemstillingen rundt variabelen «ad_count» og hvordan dette har innvirkning på både den avhengige variabelen («object_price») og den uavhengige variabelen «mileage». Konsekvensen ble at flere av datapunktene har kunstig høye verdier for disse to variablene. I datasett 3 er samtlige datapunkt som ble påvirket av denne problemstillingen fjernet. Det kan derfor tenkes at årsaken til at datasett 3 gir et lite positivt utslag i forhold til datasett 1, er at det har en mer «naturlig» fordeling og variasjon i dataen. Det er riktignok snakk om veldig liten forskjell, og vi kan derfor ikke si dette med sikkerhet.

Datasett 1:

	LR	**DTR**	**NN**
R^2	0,8536	0,8299	0,8977
MSE	800 755 100	930 518 714	559 638 145
RMSE	28 298	30 504	23 657
MAE	21 722	21 202	17 428
MAPE	0,1641	0,1467	0,1189

Tabell 5.5: Evalueringsstatistikk for modellene med datasett 1.

Datasett 3:

	LR	**DTR**	**NN**
R^2	0,8547	0,8446	0,8994
MSE	800 116 600	857 338 009	554 216 917
RMSE	28 286	29 280	23 542
MAE	21 687	20 448	17 456
MAPE	0,1640	0,1385	0,1203

Tabell 5.6: Evalueringsstatistikk for modellene med datasett 3.

5.4.2 Modell

Etter å ha evaluert hver modelltype for seg, vil vi nå drøfte hvilken av modellene som evner å tilpasse seg dataen best, og gi de mest presise prediksjonene.

Basert kun på statistiske måltall, er det decision tree regresjon som utpeker seg som den dårligste modellen, imens det er nevrale nettverk som gir de beste resultatene. Lineær regresjon er tilsynelatende litt bedre enn decision tree regresjon. Analysene i de foregående

kapitlene har derimot avdekket at disse måltallene ikke alltid gir en fullstendig og god representasjon av modellenes prestasjoner og tilpasninger.

Analysen av spredningsplott og prisfordelinger avslørte at de lineære regresjonsmodellene har flere svakheter. Modellene predikerer for lavt i noen prissjikter, men samtidig for høyt i andre prissjikter. Det er også tilfeller av negative prediksjonsverdier. Jevnt over gjør dette at modellen ikke treffer spesielt bra i noen prisintervaller, og beviser at lineær regresjon ikke evner å fange opp kompleksiteten i dataen.

Decision tree regresjon har klart og tydelig størst problemer med nøyaktigheten av alle modellene, og spredningsplottene viser flere store avvik. Det som likevel kan trekkes fram som positivt i forhold til de lineære regresjonsmodellene er at avvikene har en vifteform; størrelsen på avvikene stiger med økende salgspriser. Det betyr at modellen har tydelig bedre treffsikkerhet for de lavere prissjiktene. Der de lineære regresjonsmodellene bommer såpass mye at det ikke er mulig å finne noen intuitive bruksområder for dem, presterer decision tree regresjon såpass bra for de billigste bilene at den er en reell kandidat for videre analyser. Dette, til tross for at decision tree regresjon ikke så veldig lovende ut kun basert på de statistiske måltallene.

Modellene utviklet med nevrale nettverk gir de klart beste resultatene, både sett i betraktning av de statistiske måltallene, og de ulike grafiske framstillingene av prediksjonsresultatene. Dette er ikke veldig overraskende med tanke på modellens kompleksitet; den har større potensiale for å fange opp mønstrene i datasettet og finne den beste tilpasningen. Til tross for at modellene med nevrale nettverk har minst avvik, har de likevel noen tydelige svakheter. Modellene (NN1 og NN3) er for det meste veldig treffsikre, men bommer til gjengjeld kraftig når de først bommer. Dette er spesielt et stort problem i det høyeste prissjiktet hvor modellen rett og slett ikke evner å predikere verdier over et visst prisnivå. Det er også lignende tendenser i det laveste prissjiktet, men her kommer ikke modellens avgrensning like tydelig fram.

På bakgrunn av disse funnene kan det argumenteres for at både decision tree regresjon og nevrale nettverk er gode kandidater som prismodeller for bruktbilsalg av Nissan Leaf. Det må riktignok påpekes at de nevrale nettverksmodellene presterte tydelig bedre i disse analysene, og at det også er disse modellene som har best muligheter for forbedring ved videre arbeid.

6. Konklusjon

Oppgavens formål var å finne den beste prediksjonsmodellen for bruktbilsalg av Nissan Leaf. For å avgrense analysen ble det valgt tre ulike modeller; lineær regresjon, desicion tree regresjon og nevrale nettverk. På grunn av problemstillinger tilknyttet datagrunnlaget, ble det laget tre ulike versjoner av datasettet. Hensikten bak dette var både å sikre kontroll over mulige feilkilder, samt å få undersøkt om å skille ut gamle modeller av Nissan Leaf kan øke presisjonen til modellene.

Analysene ga både overraskende, og forventede funn. Lineær regresjon er ikke en god tilpasning for salgsprisen. Modellen er for simpel, og det kommer tydelig fram at en lineær tilpasning ikke er tilstrekkelig for å representere kompleksiteten i dataen. Dette til tross for at måltallene ga gode indikasjoner innledningsvis.

Modellene med decision tree regresjon, som opprinnelig ga dårligere resultater i form av statistiske måltall, predikerer godt i det laveste prissjiktet. Prediksjonene blir tydelig mindre treffsikre etter hvert som salgsprisen øker, og det er stor usikkerhet i prediksjonene av priser over kr 200 000. Selv om modellen i sum har store avvik, mener vi at den har potensiale for bruk til prediksjon av salgspriser, men da avgrenset til et bestemt prisintervall i det lavere prissjiktet.

Nevrale nettverk utpeker seg som den beste modellen, med gode, presise prediksjoner i et relativt stort prisintervall. Modellen har imidlertid en stor svakhet i form av terskelen som konstrueres for det høyeste prissjiktet. Avgrenser man bruken av modellen opptil kr 300 000 har man likevel en prediksjonsmodell som predikerer godt for et stort prisintervall.

Modellene basert på datasett 1 og 3, presterer bedre enn modeller med datasett 2. Drøftingen avslørte at prisene for de nyeste bilene har en annen fordeling enn den totale mengden salgspriser. Dette skyldes at disse bilene i snitt selges dyrere enn de eldre bilene, og er derfor jevnere fordelt med en høyere snittpris. Det er derfor noe overraskende at modellene med datasett 2 ikke evner å predikere bedre i det øvre prissjiktet enn de øvrige modellene, ettersom dataen har en større andel biler med høye priser. På grunn av at vi ikke ser forbedring med datasett 2 i forhold til de andre datasettene er det rimelig å anta utfordringene med det øverste prissjiktet skyldes at det er for lite datagrunnlag på de dyreste salgsprisene. Dette er en problemstilling som ikke løses ved å skille disse datapunktene fra den resterende dataen.

Det er små, men konsekvente forskjeller i prestasjonene til modellene med datasett 3 og 1. Funnene i denne oppgaven er ikke tilstrekkelige for å trekke en endelig konklusjon, men det peker i retning av at det har hatt en positiv effekt å fjerne datapunktene som måtte bearbeides i forkant av analysen.

For eventuelt videre arbeid er det flere ting man kan gjøre for å forsøke å utbedre prediksjonsevnene til modellene med tanke på prising av elektriske bruktbiler. Det første vil være å inkludere flere elektriske bilmodeller, og gjerne biler i en høyere prisklasse da våre modeller predikerte dårlig i det segmentet. Det hadde også vært gunstig å få inn flere variabler som størrelse på bil, farge, mer konkrete salgstidspunkter og generelt mer datapunkter som modellen kan basere seg på. Til slutt så kunne man forsøkt å anskaffe et datasett med de endelige salgsprisene og ikke prisene fra lukkede salgsannonser slik vi har brukt her. Det kunden betaler til slutt kan variere fra prisen i salgsannonsen som følge av pruting mellom kjøper og selger.

Referanser

Abrahamsen, M. (2024) Bruktbiler: Gammel traver størst på Finn, *Motor*. Tilgjengelig fra: https://www.motor.no/aktuelt/nissan-leaf-storst-pa-finn-i-2023/262752 (Hentet 14. februar 2024)

Aggarwal, C.C. (2018) *Neural Networks and Deep Learning.* Sveits: Springer. Tilgjengelig fra: https://link.springer.com/book/10.1007/978-3-319-94463-0

Ahite, J.B. (2018) *The Artificial Neural Networks handbook: Part 1*. Tilgjengelig fra: https://www.datasciencecentral.com/the-artificial-neural-networks-handbook-part-1/ (Hentet: 20. april 2024).

Breiman, L. (1984) *Classification and Regression Trees*. University of Michigan: Wadsworth International Group.

Ege, T. (2024) *Ny lovendring gir deg bedre forbrukerrettigheter*. Tilgjengelig fra: https://www.forbrukerradet.no/siste-nytt/ny-lovendring-gir-deg-bedre-forbrukerrettigheter/ (Hentet 15. april 2024)

Hastie, T., James, G., Tibshirani, R. og Witten, D. (2013) *An Introduction to Statistical Learning with Applications in R.* 1. utg. New York, Heidelberg, Dordrecht, London: Springer Tilgjengelig fra: https://www.stat.berkeley.edu/users/rabbee/s154/ISLR_First_Printing.pdf

Hastie, T., Tibshirani, R., og Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2. utg. New York, Heidelberg, Dordrecht, London: Springer.

Hyndman, R. J. og Athanasoupolus, G. (2021) *Forecasting: Principles and Practice*. 3. utg. Monash University: Otexts. Tilgjengelig fra: https://otexts.com/fpp3/accuracy.html

Johnson, K. og Kuhn, M. (2013) *Applied Predictive Modelling*. 1. utg. New York, Heidelberg, Dordrecht, London: Springer. Tilgjengelig fra: https://vuquangnguyen2016.files.wordpress.com/2018/03/applied-predictive-modeling-max-kuhn-kjell-johnson_1518.pdf

Montgomery, D.C., Peck, E.A., og Vining, G.G. (2012) *Introduction to Linear Regression Analysis.* 5. utg. Hoboken, New Jersey: Wiley.

Nissan Motor Corporation (2020) *A decade of innovation – the LEAF's incredible journey.* Tilgjengelig fra: https://www.nissan-global.com/EN/STORIES/RELEASES/nissan-leaf-10years/ (Hentet: 20. Februar 2024)

Pullen, J. og Sanderson, G. (2024a) *But what is a Neural Network?* Tilgjengelig fra: https://www.3blue1brown.com/lessons/neural-networks (Hentet: 20. april 2024)

Pullen, J. og Sanderson, G. (2024b) *Gradient descent, how neural networks learn.* Tilgjengelig fra: https://www.3blue1brown.com/lessons/gradient-descent (Hentet: 22. april 2024)

Quinlan, J. R. (1986) *Induction of Decision Trees. Machine Learning*. Boston: Kluwer Academic Publishers.

Sørensen, A. (2023) 2023 kan bli enda dyrere: - Jeg må være forsiktig, *NRK*. Tilgjengelig fra: https://www.nrk.no/mr/nyttarsforsett-for-2023-_-folk-er-mer-bevisst-okonomisk-og-vil-spare-1.16237179 (Hentet 14. januar 2024)

Whitfield, B. (2022) *Feedforward Neural Networks: A Quick Primer for Deep Learning* Tilgjengelig fra: https://builtin.com/data-science/feedforward-neural-network-intro (Hentet: 22. april 2024)

Vedlegg

Vedlegg 1

Variabler	**Forklaring**
county	Kategorisk variabel for hvilken landsdel bilen har blitt annonsert i; Midt-Norge, Nord-Norge, Sørlandet, Vestlandet og Østlandet.
sales_channel	Kategorisk variabel angående selgerens natur; privatpersoner (Privat), bruktbilforhandlere (Annen bilutsalg) og merkeforhandlere.
year_model	Registreringsåret til bilmodellen det er snakk om. Inkluderer 2013-modeller til 2023-modeller.
sales_year	Året bilen ble solgt på Finn.no.
ad_count	Antall ganger dette skiltnummeret har blitt annonsert av samme aktør.
alive_days	Antall dager annonsen lå ute på Finn.no før den ble lukket.
object_price	Prisen i salgsannonsen.
mileage	Kilometerstanden på bilen.

Vedlegg 2

county:

- 0: Nord-Norge
- 1: Midt-Norge
- 2: Sørlandet
- 3: Vestlandet
- 4: Østlandet

➔ Rangering basert på avstander og folketall.

sales_channel:

- 0: Privat
- 1: Annet bilutsalg
- 2: Merkeforhandler

➔ Rangering basert på økende grad av profesjonalitet.

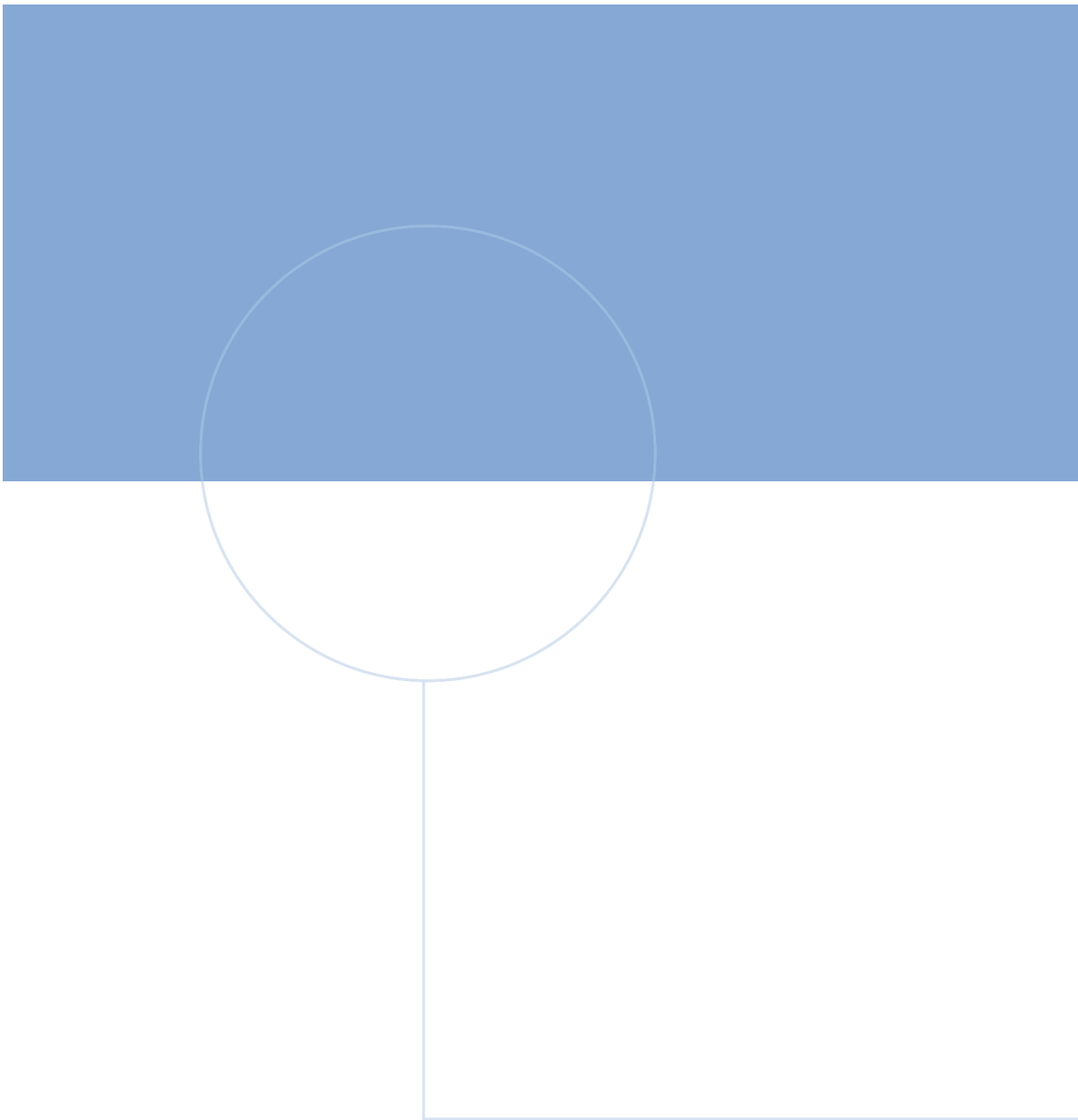
Vedlegg 3

Bibliotek	
pandas	Funksjoner for databehandling.
matplotlib.pyplot	Funksjoner for visuelle presentasjoner av data.
sklearn .model_selection .preprocessing .metrics .linear_model	Funksjoner for databehandling før analyser, funksjoner for regresjonsanalyser, samt funksjoner for å beregne statistiske måltall: train_test_split, min_max_scaler, r2_score, Label_encoder, mean_squared_error, mean_absolute_error, mean_absolute_perecentage_error, LinearRegression, DecisionTreeRegressor.
keras .models .layers .callbacks	Funksjoner nevrale nettverk; Sequential, Dense, InputLayer, EarlyStopping.
numpy	Funksjoner for matematiske beregninger.

Vedlegg 4 – Kode

For kode tilhørende oppgaven; se vedlagte pdf-filer.

Datasettene brukt i analysen er konfidensielle, og vil derfor ikke bli gjort tilgjengelig.



NTNU

Kunnskap for en bedre verden