Bacheloroppgave

NTNU
Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

Andreas Kjosavik
Markus Neupauer Presthus
Morten Selbekk Husby
Silje Marie Østberg Mogen

Modell for prediksjon av eiendomspriser i Oslo

Bacheloroppgave i Økonomi og administrasjon
Veileder: Mike Denis Becker

April 2022

NTNU
Kunnskap for en bedre verden

Andreas Kjosavik
Markus Neupauer Presthus
Morten Selbekk Husby
Silje Marie Østberg Mogen

Modell for prediksjon av eiendomspriser i Oslo

Bacheloroppgave i Økonomi og administrasjon
Veileder: Mike Denis Becker
April 2022

Norges teknisk-naturvitenskapelige universitet
Fakultet for økonomi
NTNU Handelshøyskolen

NTNU
Kunnskap for en bedre verden

Forord

Denne bacheloroppgaven markerer avslutningen av vårt treårige studieforløp ved NTNU Handelshøyskolen. Gjennom arbeidet med denne oppgaven har vi tilegnet oss dypere forståelse for blant annet boligprispredikering, hedonisk prismodell og kunstig nevrale nettverk. Prosessen med å skrive denne oppgaven har vært lærerikt, interessant og til tider utfordrende. Innholdet i denne oppgaven står for forfatteren(e)s regning.

Vi vil benytte anledningen til å takke vår veileder, Denis Becker, for godt samarbeid og god veiledning i løpet av våren. Vi setter stor pris på den faglige støtten, engasjementet og tilgjengeligheten gjennom bacheloroppgavens forløp. Vi vil også takke solgt.no for å dele eiendomsdata med oss. Uten denne dataen hadde det ikke vært mulig å få gjennomført oppgaven slik vi ville.

Sammendrag

Formålet med denne oppgaven var å bruke maskinlæring til å lage en modell for å kunne predikere eiendomspriser i Oslo. En typisk modell å bruke for prispredikering er hedonisk prismodell. Med bakgrunn i dette har vi også valgt å se på hedonisk prismodell, både som sammenligningsgrunnlag og for å bedre forstå hvordan prispredikering vanligvis foregår. For å utvikle modellene våre har vi brukt nevrale nettverk og hedonisk regresjon. Datasettet vårt inneholdt mange datapunkter og vi har begrenset modellen våres en del. Dette betyr også at bruksområdene til modellene er begrenset.

Vi har utviklet to modeller med nevrale nettverk, hvor den ene har en makspris på bolig på 15 millioner kroner (ordinær modell), mens den andre har en makspris på 5 millioner kroner (modell for førstegangskjøper). Når vi begrenser maksprisen til 5 millioner kroner ser vi på boliger som kanskje er mest interessante for førstegangskjøpere. Videre har vi utviklet to modeller hvor vi har brukt hedonisk metode. Den ene modellen er det brukt multippel regresjon uten noen tilpasninger, med de samme begrensningene som er benyttet i ordinær modell. I den andre modellen er det gjort en logaritmisk transformasjon av den avhengige variabelen, og dataen er begrenset på samme måte som i modell for førstegangskjøpere. Vi fikk svært lovende resultater fra modellene, hvorpå den ene modellen hadde en R^2- score opp mot 90%.

Konklusjonen fra denne oppgaven er at hvilken modell som er best, avhenger av bruksområdet som modellen benyttes til. Ønsker man en modell hvis virkemåte er enkel å forstå bør man velge en av modellene som benytter hedonisk regresjon, mens modellene som benytter nevrale nettverk gir økt presisjon. Er presisjon særlig viktig, bør man velge en av modellene med utvidete begrensninger i dataene, og vice versa.

Abstract

The purpose of this thesis is to utilize machine learning to make a model that could predict real estate prices in Oslo. A typical model to use for price prediction is hedonic price model. With this in mind we have also chosen to look at hedonic price model, both as basis for comparison and to get a better understanding of how price prediction usually takes place. To develop our models, we have used neural networks and hedonic regression. Our dataset contained many data points and based on this we limited our models quite a bit. This also means that the models cannot be used in a lot of areas.

We have developed two models with neural networks, where one of them have a maximum price on housing of 15 million NOK (ordinary model), while the other have a maximum price of 5 million NOK (model for first time buyers). When we limit the maximum price to 5 million NOK, we look at housing that may be most interesting for first time buyers. Furthermore, we have developed two models using hedonic price model. On the first model we have used multiple regression without any adjustments, but with the same limitations used in the ordinary model. In the second model there is done a logarithmic transformation of the dependent variable, and the data is limited in the same way as the model for first time buyers. We got very promising results from our models, where one of them had an R^2-score closing to 90%.

The conclusion from this thesis is that the models that performs the best, depends on the area of use. If you want a model that are easy to use, you should use one of the models that are based on hedonic price model. The models using artificial neural network gives better precision. If precision is especially important, you would choose one of the models with extended restrictions, and vice versa.

Innholdsfortegnelse

Innholdsliste figurer

1.0 Innledning

For nordmenn er det å eie en egen bolig viktig, noe som også gjenspeiler seg i den norske økonomien. I 2020 eide 81,8% av nordmenn sin egen bolig og i 2019 utgjorde formuen til nordmenn tilknyttet boliger 8 528 milliarder kroner (SSB, 2018; Eiendomnorge, 2019). Det levner liten tvil om viktigheten for norsk økonomi i å ha et best mulig beslutningsgrunnlag for å gjennomføre investeringer i boliger. I den forbindelse skal vi lage en automatisk verdsettelsesmodell for boliger i Oslo. En slik modell kan være av interesse for alle aktører som befatter seg med eiendomsmarkedet. For eksempel vil banker være interessert i en mest mulig presis verdivurdering når de skal håndtere risiko i sine utlånsporteføljer, hvorav en stor andel vil være sikret i nettopp boliger. I tillegg vil Skatteetaten være interessert i en presis verdivurdering for å beregne likningsverdi på bolig og formuesskatt.

Som nevnt skal vi lage en automatisk verdsettelsesmodell for boliger. I den sammenheng vil vi benytte to teknikker: hedonisk prismodell og kunstig nevralt nettverk. Metodene har hver sine styrker og svakheter, og vil som sådan egne seg for ulike bruksområder. Vi skal i diskusjonsdelen se nærmere på dette. For å lage en modell som tilfredsstiller minimumskrav til presisjon har vi valgt å gjennomføre betydelige avgrensninger i modellens omfang. I første omgang dreier dette seg om å ta for seg mest mulig ordinære boliger. Dette vil vi komme tilbake til i seksjonen om data. Som følger vil modellene ha et begrenset gyldighetsområde, for eksempel er den uegnet til å ta for seg næringseiendom.

Strukturen i oppgaven vil være som følgende: først vil vi foreta en litteraturgjennomgang av tidligere studier innen automatisk verdsetting. Deretter vil vi ta for oss teori om det norske boligmarkedet, med særlig fokus på Oslo. Herunder vil også den hedoniske prismodellen, maskinlæring og nevrale nettverk redegjøres for. I dataseksjonen presenteres hvilke begrensninger vi har gjort for å kunne lage modellen og en redegjøring av datarensning. Metodeseksjonen tar for seg fremgangsmåten og hvilke valg som er gjort i forbindelse med utvikling av modellene. Videre vil vi i seksjonen for empiriske resultater ta for oss resultatene fra modellene og til slutt diskutere egenskaper ved de ulike modellene.

2.0 Litteraturgjennomgang

Denne delen presenterer relevant litteratur til boligprispredikering. En analyse av boligmarkeds- og boligverdsettelseslitteratur viser to hovedtrender: hedonisk regresjonstilnærming og maskinlæringsmetoder for å utvikle boligprisprediksjonsmodeller.

Studien av Rosen (1974) ga en detaljert forklaring av HPM (hedonisk prismodell). Etter studien har forskjellige eiendomsmarkeder rundt om i verden blitt modellert ved å bruke HPM for å måle bidragskraften av forskjellige klassifikasjoner (lokasjon, nabolag og det strukturelle) til fastsetting av eiendomsverdier. I flere tiår ble HPM brukt for å indentifisere sammenheng mellom boligpriser og boligkarakteristikker. I følge Abidoye & Chan (2018), ble ANN-modellen først brukt i eiendomsvurdering av Borst i 1991. Funnene fra studien viste at ANN-modellen kunne produsere pålitelige og nøyaktige verdivurderingsestimater. Dette førte til stor aksept av teknikken og blitt brukt i modellering av eiendomspriser i land som blant annet USA, Irland, Spania og Italia.

HPM har i midlertidig mulige begrensninger knyttet til grunnleggende modellforutsetninger og estimering. Forskere har hevdet at ANN-teknikken ble tilpasset eiendomsvurdering for å adressere mangler ved HPM. Nyere studier har derfor fokusert på sammenligning av prisprediksjonsytelsen mellom hedonisk-basert modeller og maskinlæringsalgoritmer (Bae & Park, 2015).

Selim (2009) sammenlignet prediksjonsytelsen mellom HPM og ANN-modeller. Denne studien viste at ANN-modellen kan være et forbedret alternativ for prediksjon av boligpriser i Tyrkia. Abidoye og Chan (2018) evaluerer den prediktive nøyaktigheten til hedonisk modell sammenlignet med nevrale nettverk i eiendomsvurdering. Eiendomsdataen ble samlet inn fra eiendomsfirmaer i Lagos, Nigeria. Resultatene viste at ANN-modellen overgikk den hedoniske modellen, når det gjelder nøyaktigheten i å predikere eiendomsverdier. Buodd og Derås (2022) undersøkte om maskinlæringsmetoder kan forbedre eiendomsprispredikeringen, som igjen fører til mer nøyaktig eiendomsskatteanslag i Norge. De undersøkte tre distrikter i Oslo og brukte flere metoder, blant annet nevrale nettverk. Resultatene viste at alle metodene, utenom beslutningstre, presterte bedre enn multippel lineær regresjon. Hovedkonklusjonen var at maskinlæringsmetoder var i stand til å forbedre dagens metoder for eiendomsskatteanslag i Norge.

3.0 Teori

3.1 Boligmarkedet

3.1.1 Boligmarkedet i Norge

Det norske boligmarkedet kjennetegnes ved en høy eierandel hvor de fleste husholdningene eier og bor i sin egen bolig, og blir sett på som den viktigste investeringen for de fleste norske husholdninger. Boligmarkedets utvikling kan ha betydelige konsekvenser for konsumentene og økonomien i helhet. Sentralt for en velfungerende økonomi er et sunt boligmarked (Walbækken & Bendictow, 2020).

Den mest brukte digitale plattformen ved kjøp og salg av eiendom i Norge er finn.no. Når man publiserer en eiendomsannonse, plasseres den øverst i listen. Etter hvert som flere annonser legges ut, vil den rykke nedover i listen. Finn tilbyr ulike pakker slik at boligannonsen får en bedre effekt, og det blir lettere å selge boligen. Eiendomsmegleren må skape oppmerksomhet og tillit slik at leseren får en følelse av at det er trygt å investere pengene sine, samt er det avgjørende at leserne finner den informasjonen de er ute etter (Finn.no, 2022).

3.1.2 Boligmarkedet i Oslo

Boligmarkedet i Oslo kjennetegnes ved å ha de høyeste og voksende boligprisene i Norge. Det er et klart skille mellom øst og vest når det gjelder prisnivået på boliger. Dette skillet kommer historisk fra økonomiske og sosiale forhold. Arbeiderklassen bodde på østkanten og borgerskapet bodde på vestkanten. I senere tid har mange innvandrere utenfor Europa bosatt seg på østkanten, noe som har bidratt til å opprettholde dette skillet (Barlindhaug, 2005). For eksempel så kan noen boliger i de sørøstlige delene av Oslo være priset halvparten av hva boligene i de mest attraktive områdene i Oslo indre vest omsettes for. Gjennomsnittlig kvadratmeterpris for blokkleiligheter i Frogner per. 2019 var 90 332 kroner og 43 210 kroner i Søndre Nordstrand (Bydelsfakta, 2019). I de siste årene har dette skillet blitt mindre pga. de høye prisene i sentrum, noe som har ført til at folk i større grad trekker seg østover (Barlindhaug, 2005).

Oslo kan deles inn i 16 bydeler, inkludert Oslo sentrum. En oversikt over dette vises i figur 1. Oslo består av 76% leiligheter, blokker ol. For de sentrumsnære bydelene som Sagene, St. Hanshaugen, Gamle Oslo og Grünerløkka, så ligger leilighetsandelen over 90%. Det er i

motsetning til bydelene lenger borte fra sentrum, som Nordstrand og Vestre Aker, hvor leilighetsandelen er rundt 40% (Byrådsavdeling for finans, 2021).

Figur 1: Grov inndeling av Oslos bydeler (Lokalwiki, 2004)

3.2 Hedonisk metode

Hedonisk metode er en prismodell som baserer seg på at prisen på et gode i et gitt marked reflekteres av en samling av ulike attributter (egenskaper). Observerte priser på attributtene og mengden av egenskaper assosiert med dem danner et sett med implisitte priser. Videre antar Rosen at forbrukere vil velge de godene som inneholder den samling av attributter som vil maksimere deres nyttefunksjon (Rosen, 1974). Hedonisk pristeori stammer fra Lancasters (1966) forbrukerteori hvor han tar utgangspunkt i at det er egenskapene ved en vare og ikke varen i seg selv som gjør at forbrukeren velger den ene over den andre.

De implisitte eller "hedoniske" prisene kan observeres indirekte fra totalprisen på godet, hvor totalprisen er en funksjon av mengden attributter $Z = (Z_1, ..., Z_n)$ og deres implisitte pris. Osland (2001) definerer den implisitte prisen som økning i samlet pris på godet ved en marginal partiell økning i mengden av et attributt. Den hedoniske prisfunksjonen blir dermed $P(Z) = P(Z_1, ..., Z_n)$, hvor *n* beskrives som antall attributter under observasjon. Z er en vektor på et plan i flere dimensjoner hvor både kjøpere og selgere befinner seg, og Z_i måler mengden av egenskaper (i) i hvert attributt (Rosen, 1974). Både produsentenes tilbud og kjøperens etterspørsel er dermed viktig for prisfunksjonen (Osland, 2001).

Boligattributter deles inn i to hovedgrupper. Den første er attributter utenfor boligen, som f. eks avstandsmålinger og sosiale faktorer. Den andre hovedgruppen er attributter knyttet til selve boligen, som f. eks antall rom og antall peiser (Osland, 2001). Fordelen med hedonisk metode er at man kan kontrollere for egenskapene ved boligen, og dermed få bedre oversikt over innvirkningene en endring ved attributtene fører med seg. Metoden har likevel noen ulemper. Problemer slik som utstikkere, multikollineritet, heteroskedastisitet, uavhengige datapunkt og ikke-linearitet kan føre til at modellen ikke yter bra nok til eiendomsvurderinger (Limsombunchai, 2004).

Chin og Chau (2003) gjennomfører i sin artikkel en kritisk vurdering av litteraturen basert på hedonisk prismodell, og fremlegger en liste med attributter som vanligvis brukes i boligpredikering. Denne listen og dermed attributtene deles inn i tre deler; attributter knyttet til lokasjon, attributter knyttet til struktur og attributter knyttet til nabolaget. Strukturelle attributter vil være det vi tidligere har nevnt som attributter knyttet til selve boligen. I listen viser Chin og Chau også til hvilken effekt attributtene har for boligprisen. For eksempel viser de til at utsikt har en positiv effekt på boligprisen, mens alder på bygningen trekker boligprisen ned. For full liste se vedlegg 3.

3.3 Maskinlæring og nevrale nettverk

Maskinlæring dreier seg om å finne skjulte mønstre i et datasett og benytte det til å klassifisere eller predikere en hendelse. Det er vanlig å skille mellom to hovedformer for maskinlæring: veiledet læring og ikke-veiledet læring. Veiledet læring innebærer at maskinen får inngangsverdier som skal forutsi utgangsverdier. Ikke-veiledet læring vil derimot ikke ha utgangsverdier for inngangsverdiene. Istedenfor skal maskinen forsøke å finne en struktur i dataene, eksempelvis ved å gruppere like enheter. Nevralt nettverk er et tilfelle av veiledet læring, og ikke-veiledet maskinlæring vil derfor ikke bli videre utforsket i denne oppgaven (Buodd & Derås, 2020).

Et sentralt aspekt ved maskinlæring er at algoritmen trenes opp på data, med andre ord er det ikke regelstyrt. Etter adekvat trening skal modellen være i stand til å gjøre prediksjoner på nye observasjoner med en viss grad av presisjon. I så legger man til grunn en antakelse om at det er en sammenheng mellom forklaringsvariablene og responsvariabelen. Det kan illustreres i følgende likning:

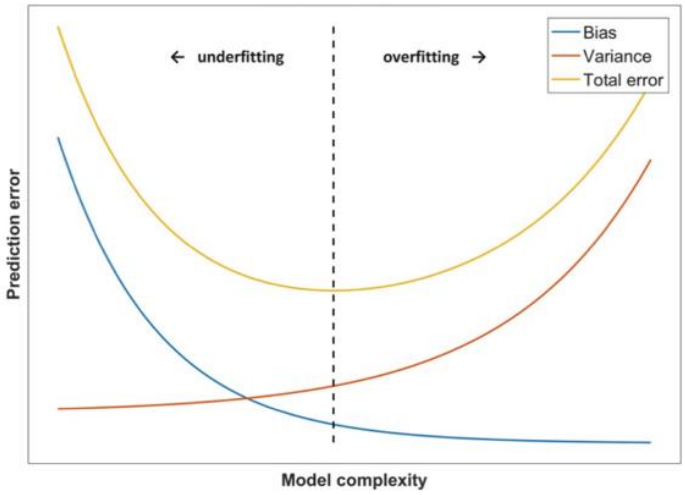
$$Y = f(x) + \epsilon$$

I likningen representerer *f* en fast, ikke-observerbar funksjon og epsilon representerer et tilfeldig feilledd, uavhengig av X og med forventning lik null. Poenget her er at det er en systematisk sammenheng mellom forklaringsvariablene X og responsvariabelen Y som man forsøker å kartlegge. Når modellen gjør prediksjoner, ønsker man å kunne bedømme disse. Til den hensikt benytter man en tapsfunksjon som ser på avviket mellom prediksjon og faktisk verdi. I videre forstand kan en slik tapsfunksjon benyttes til å sammenligne ulike modeller (Buodd & Derås, 2020).

Når man utarbeider en modell, vil målet naturligvis være å minimere avviket mellom prediksjon og faktisk verdi. For enhver observasjon x_0 kan avviket til prediksjonen forståes slik:

$$Err(x_0) = feilmargin + skjevhet^2 + varians$$

Feilmarginen er ikke noe man kan påvirke. Det skyldes variasjonen rundt det sanne gjennomsnittet for hendelsen man søker å predikere utfallet av. Som eksempel kan to eksakt like boliger bli solgt for ulike priser, men en modell vil ikke evne å fange opp dette. Skjevhet sier noe om avviket mellom modellens gjennomsnitt for estimatet og det sanne gjennomsnittet for denne variabelen. Det siste leddet, varians, dreier seg om forventet kvadratavvik mellom prediksjon og faktisk verdi (Hastie et. al., s. 37). For å utdype: en modell med *høy* varians som trenes på nye data anskaffet under like rammer vil gi vidt forskjellige prediksjoner, med andre ord vil det være en lite robust modell. Mens feilmarginen er utenfor ens kontroll, er skjevhet og varians noe man må ta høyde for når man utarbeider en modell. Egenskapene til disse to leddene gir opphav til en avveiing, og er noe man søker å optimere i prosessen med å lage modellen. Denne avveiingen er illustrert i figur 2.



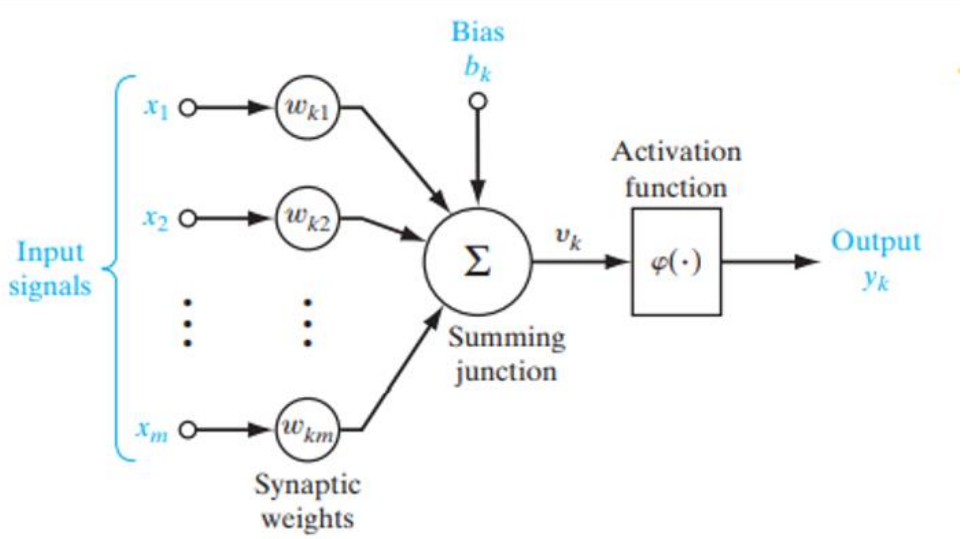
Figur 2: Prediksjonsfeil. (Buodd & Derås, 2020).

Det kommer frem tre sentrale begreper av denne figuren: undertilpasning, overtilpasning og modellkompleksitet. Undertilpasning vil gjelde for en modell som er for enkel, og som dermed ikke evner å fange opp de underliggende mønstrene på en tilfredsstillende måte. På den andre siden vil overtilpasning være tilknyttet en for kompleks modell, som vil være utsatt for støy og egenskaper ved det konkrete datasettet (Hastie et. al., s. 38). Spørsmålet blir så hva som bestemmer kompleksiteten til en modell. Som regel vil det handle om antall parametere i modellen; dess flere parametere, jo mer kompleks er modellen. Det er igjen knyttet til hyperparametere, som kan forståes som parametere hvis verdi er styrende for treningen av en modell. Med andre ord er det tett tilknyttet valgene vedkommende som skal konstruere en modell tar, i tillegg vil dataene man benytter ha betydning (Buodd & Derås, 2020).

Et nevralt nettverk er en datastruktur som er laget etter inspirasjon av hjernens grunnleggende virkemåte (Dvergsdal, 2019). Nevrale nettverk har vist seg å være godt egnet innenfor mange bruksområder som for eksempel bilde-, tale- og tekstgjenkjenning, og regnes som en universal approksimeringsmetode (Hornik, Stinchcombe & White, 1989). Ovenfor ble det redegjort for at man i maskinlæring søker å kartlegge en ukjent funksjon mellom forklaringsvariablene og responsvariabelen. At nevrale nettverk regnes som en universal approksimeringsmetode innebærer at den er god på å estimere gitt funksjon, for alle ulike varianter av slike funksjoner.

Strukturen til et nevralt nettverk består av inngangslaget, utgangslaget og skjulte lag. Hvert lag består av noder som hver for seg er koblet til alle nodene i det neste laget via synapser

med vekting. Synapsevektene bestemmer styrken i signalet fra en node til den neste. En node vil summere verdiene den mottar fra andre noder, legge til en bias og deretter sende dette tallet inn i en aktiveringsfunksjon før det sendes inn i synapsene. Det kan i teorien være uendelig antall noder og lag i et nevralt nettverk og det er et stort utvalg med aktiveringsfunksjoner man kan benytte, eksempelvis sigmoid og logistisk. Synapsevektene og biasene er det som blir bestemt gjennom trening av modellen. Antall lag og noder, samt aktiveringsfunksjoner, i hvert lag er hyperparametere. Se figur 3 for en illustrasjon av oppbygningen. (Dvergsdal, 2019).



Figur 3: Oppbygning av en enkelt node. (Buodd & Derås, 2020).

4.0 Data

Datasettet vi har benyttet i denne oppgaven er skaffet i samarbeid med solgt.no. Det er et datasett med informasjon om eiendommer i Oslo som er omsatt mellom 2008 og 2022. Før vi startet med rensingen av dataen så inneholdt settet informasjon om 181 250 forskjellige eiendommer, med totalt 57 forskjellige variabler. Etter vi hadde renset dataen til ordinær modell sto vi igjen med 177 407 datapunkter, mens vi etter rensing til modell for førstegangskjøpere sto igjen med 102 764 datapunkter. Se vedlagt Jupyter-fil for en mer detaljert gjennomgang av dataen.

For at datasettet skal kunne brukes til analyse må man rense dataen for å håndtere null-verdier og andre feilverdier. Det første vi gjorde var å gå gjennom hver variabel og så på fordelingen, antall null-verdier og eventuelle feilverdier. Deretter fant vi en løsning for hver enkelt variabel, alt etter hva som virker best. For modellene våre vil det kunne være flere måter å rense dataen på, og man bør prøve seg litt frem til man får en modell som fungerer best mulig.

Vi brukte flere forskjellige teknikker for å håndtere null-verdier. For variablene som omhandlet lengde- og breddegrader så valgte vi å erstatte null-verdiene med en verdi som var sentrert i bydelen. Dermed fikk vi en verdi som samsvarer godt med hvor i Oslo boligen ligger, selv om den ikke er helt presis. Videre valgte vi å gjøre en regresjonsanalyse for å finne sammenhengen mellom PROM og BRA, slik at vi kunne erstatte null-verdier av BRA med PROM * koeffisient + skjæringspunkt. For variabelen som går på antall soverom så valgte vi først å fjerne helt klare feilverdier før vi erstattet null-verdiene med et gjennomsnitt. Her prøvde vi også å se om dette kunne kobles mot PROM/BRA, men dette hadde lite korrelasjon, så vi valgte å bruke gjennomsnitt.

For variabelen som omhandler etasje så var en del av null-verdiene fra boliger som enebolig og rekkehus hvor dette ikke er relevant. For de resterende null-verdiene så valgte vi å erstatte dem med gjennomsnittet, samt legge til en dummy-variabel som sier om variabelen var null-verdi eller ikke. For fellesgjeld og felleskostnader så valgte vi å erstatte null-verdier med 0, da det er mer sannsynlig at disse boligene faktisk ikke har disse kostnadene enn at det er null-verdier. I tillegg la vi til en indikator som sier om vi hadde erstattet en null-verdi. For alle verdiene som omhandlet finn-fasiliteter så valgte vi å erstatte null-verdier med 0 da disse kun

er 1 hvis fasiliteten finnes. For en mer detaljert oversikt over datarensingen vises det til Jupyter-filen(e).

5.0 Metode

I dette kapittelet presenterer vi fremgangsmåten og metoden brukt for å utvikle våre modeller. For både hedonisk prismodell og nevralt nettverk har vi først tatt utgangspunkt i en "enkel" ordinær modell, og deretter begrenset dataen mer i modell for førstegangskjøpere, som fokuserer på rimeligere boliger.

På bakgrunn av at vi ønsker å se på "vanlige" boliger i Oslo velger vi å begrense dataen vår på flere felt. Det å lage en modell som kan predikere prisen på alle boliger vil mest sannsynlig være umulig, og grunnet dette må vi avgrense dataen for å få en mer presis modell. Vi har prøvd med flere forskjellige avgrensninger for å se hvor bra modell vi klarer å lage. Den første modellen vi lagde, ordinær modell, begrenset vi med følgende kriterier:

- Fjerner boliger bygget før 1850.
- Fjerner leiligheter med primærrom større en 250 m^2, og eneboliger, rekkehus og tomannsboliger med primærrom større enn 350 m^2.
- Fjerner boliger med pris over 15 millioner kroner.

I modellen for førstegangskjøpere har vi begrenset dataen mer slik at vi retter boligene mot en typisk førstegangskjøper. Vi hadde da følgende begrensinger (i tillegg til begrensingene fra ordinær modell):

- Fjerner boliger som ikke er leiligheter.
- Fjerner boliger med BRA > 75 kvm.
- Fjerner boliger med pris over 5 millioner kroner.

Dette gjorde at vi fjernet 76 885 datapunkter slik at vi sto igjen med 102 764 datapunkter.

I teoridelen ble det nevnt at man har en tapsfunksjon for å vurdere kvaliteten til prediksjonene. I den forbindelse er målet å minimere gjennomsnittlig avvik mellom prediksjonene og faktisk verdi. Som tapsfunksjon vil vi benytte gjennomsnittlig kvadratisk avvik (MSE) fordi den er enkel og intuitiv, dessuten er det den mest brukte tapsfunksjonen (Pandit og Schuller, 2019). Imidlertid vil roten til gjennomsnittlig kvadratisk avvik (RMSE), gjennomsnittlig absolutt avvik (MAE) og gjennomsnittlig absolutt prosentavvik (MAPE) benyttes i prosessen med å sammenligne og diskutere modellene.

Determinasjonskoeffisienten, R^2, benyttes for å si noe om i hvilken grad modellene passer til dataene. Den matematiske definisjonen av disse er som følger:

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

$$RMSE = \sqrt{MSE}$$

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

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

$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}}$, hvor

$SS_{TOT} = \sum_{i=1}^{N}(y_i - \bar{y})^2$ og $SS_{RES} = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$.

$N = antall\ observasjoner, y = faktisk\ verdi, \hat{y} = predikert\ verdi,$
$\bar{y} = gjennomsnittlig\ verdi$

5.1 Modell med kunstig nevralt nettverk

For å lage en best mulig modell, prøvde vi oss litt frem med ulike kombinasjoner av ulike lag med ulikt antall noder. Vi har begrenset oss til å se på følgende hyperparametere: antall lag, antall noder i hvert lag, aktiveringsfunksjon, optimaliseringsalgoritme, antall gjennomkjøringer (epoch) og tidlig stopp (early stopping). For begge de kunstige nevrale nettverkene har vi benyttet optimaliseringsalgoritmen Adam. Det er den mest populære optimaliseringsalgoritmen på grunn av dens ytelse og effektivitet sammenlignet med andre algoritmer (Bock et. al., 2018). Den matematiske definisjonen av Adam er som følger:

$$For\ hver\ parameter\ w^j:$$

$$v_t = \beta_1 * v_{t-1} - (1 - \beta_1) * g_t$$

$$s_t = \beta_2 * s_{t-1} - (1 - \beta_2) * g_t^2$$

$$\Delta w_t = -\eta \frac{v_t}{\sqrt{s_t + \epsilon}} * g_t$$

$$w_t + 1 = w_t + \Delta w_t$$

$\eta: læringsrate, \qquad g_t: stigning\ på\ tidspunkt\ t\ langs\ w^j,$
$v_t: Eksponensielt\ gjennomsnitt\ av\ stigning\ langs\ w^j,$
$s_t: Eksponensielt\ gjennomsnitt\ av\ roten\ av\ stigning\ langs\ w^t,$
$\beta_1 \& \beta_2: Hyperparametre$

Vanligvis vil et større nettverk gi bedre resultater, men et større nettverk krever også mer maskinkraft for å kunne trenes. I tillegg vil dype nevrale nettverk kreve mye mer av dataene. Som tommelfingerregel skal antall observasjoner i treningssettet være minst ti ganger større enn antall parametere i modellen, for at modellen skal være i stand til å generaliseres til nye observasjoner. Løsningen vi endte opp med var et nevralt nettverk bestående av 56 input-noder, som er alle de forskjellige variablene vi har brukt. Videre hadde vi 6 lag med følgende oppbygning: 2-2-5-5-12-128-output. Dette er et resultat av et manuelt søk, kjennetegnet ved prøving og feiling. Vi benyttet ReLU aktivering for de skjulte lagene, og lineær aktivering for output-laget. ReLU er matematisk definert slik:

$$y = \max(0, x)$$

I likningen representere y utgangsverdien, mens x representerer inngangsverdien. Fordelen med ReLU er at den er enkel å forstå og skaper en sparsommelig modell. Det resulterer i et effektivt nettverk i stand til å kjøre raskere.

For å få det nevrale nettverket til å gi best resultat så er det viktig å skalere dataen. Dette er nyttig når variablene har ulike skalaer. For eksempel vil variablene pris og antall kvadratmeter være på vidt forskjellige skalaer. Til dette brukte vi MinMaxScaler fra Scikit-learn. Den fungerer slik at alle verdiene i en variabel blir skalert mellom 0-1 slik at den laveste verdien da ender opp med 0, og alle andre verdier er skalert frem til den høyeste verdien som får 1. Ved å ikke bruke denne skaleringen så fikk vi en R^2 score som var rundt 15% lavere enn når vi brukte skalering.

5.2 Hedonisk modell

Til vår hedoniske modell har vi laget en lineær regresjonsmodell med formål om å predikere eiendomspriser. Dette har vi gjort gjennom hovedsakelig bruk av SciKit-learn-biblioteket i

Python. I en forklarende regresjonsmodell er det vanligvis viktig at variablene blant annet skal være normalfordelt, p-verdiene skal være lavere enn signifikansnivået og man skal unngå multikollinaritet for at modellen skal være gjeldende. Dette er ikke like viktig i en predikerende modell, hvor hovedprioriteten er å evaluere ytelsen til prediksjonene. Det kan likevel være nyttig å se på tester om for eksempel heteroskedastisitet for å se om vi kan forbedre modellen (Becker, 2021).

Vi startet med å lage en multippel regresjonsmodell, uten inkludering av logaritmer og skalering. For å teste modellen delte vi dataen vår inn i et testsett og et treningssett, og brukte treningssettet til å utforme den lineære regresjonsmodellen. Deretter tilpasset vi modellen. Denne fremgangsmåten brukte vi på begge modellene våre; "enkel" multippel regresjon og logaritmisk transformasjon.

$$\hat{y} = \beta_0 + \sum_{i=1}^{N} \beta_i * x_i + \epsilon$$

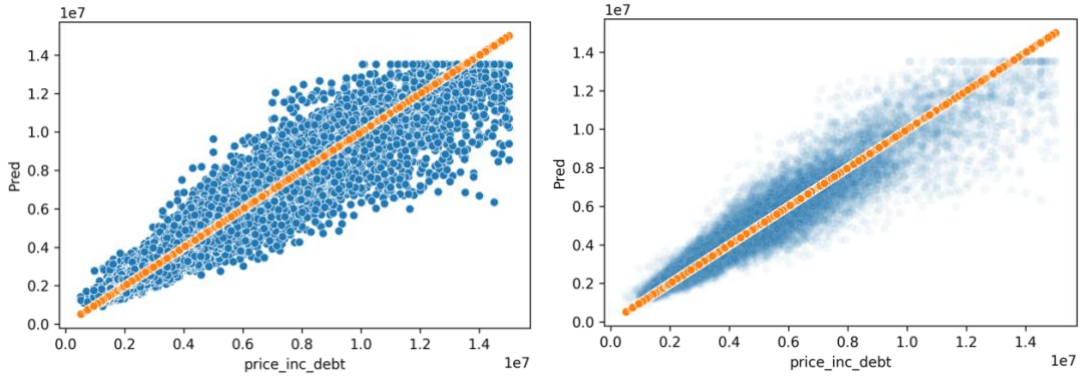
Siden vi så at den avhengige variabelen var skjevfordelt med en høyre hale, implementerte vi logaritmisk transformasjon i modellen vår. Utgangspunktet til modellen med logaritmisk transformasjon er modellen med multippel regresjon. Forskjellen i høye og lave eiendomspriser blir mindre i den logaritmiske skalaen enn i den originale målestokken i kroner, og skjevheten blir derfor redusert. Med logaritmisk transformasjon er den avhengige variabelen normalfordelt (Iversen, 2021).

6.0 Empiriske resultater og diskusjon

6.1 Nevrale nettverk

Ordinær modell:

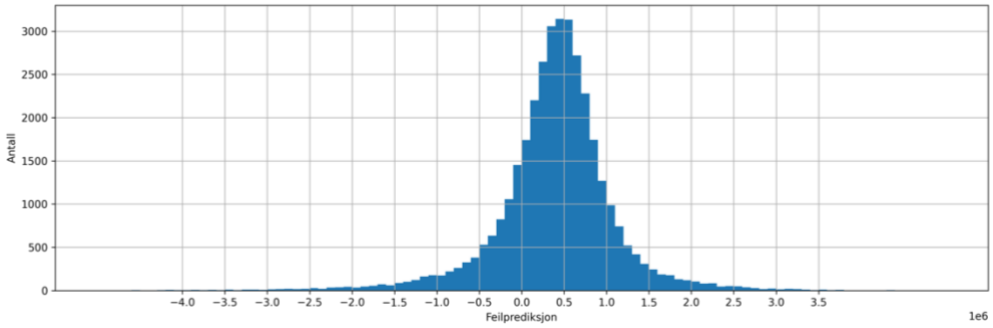
Den første modellen vi lagde, med en makspris på 15 millioner kroner, fikk en R^2 på 90,2%. RMSE for modellen er på 708 923 kroner, mens MAE er 473 490 kroner. Det vil si at modellen i betydelig grad evner å fange opp mønstrene i datasettet. Imidlertid tyder forskjellen på RMSE og MAE at det er uteliggere i avvikene mellom prediksjon og faktisk verdi. MAPE er 10,9%. Dette er resultater vi fikk etter å ha skalert dataen. Tilsvarende modell uten bruk av skalering gir R^2 på 70,84%, noe som illustrerer nytten av å bruke skalert data.



Figur 4: Feilprediksjon ANN Ordinær modell; spredningsplott

I figur 4 kan vi se to spredningsplott, med forskjellig styrke på punktene. I spredningsplottet er predikert pris plottet mot faktisk pris. De oransje punktene viser faktisk pris, mens de blå punktene modellens predikerte pris. Her vil en perfekt modell ha de blå punktene sammenfallende med de oransje punktene. I plottet til høyre ser man at brorparten av prediksjonene ligger i intervallet fra 2 millioner til 8-9 millioner kroner. Videre kan man observere at prediksjonene har en vifteform. Det vil si at modellen har mindre avvik for boliger av lavere verdi, og større avvik for boliger av høyere verdi. Det er ikke konstant varians, altså er det tilstedeværelse av heteroskedastisitet. Det korresponderer med differansen mellom RMSE og MAE. Videre kan man også observere at antallet observasjoner avtar betydelig for boliger med høyere verdi. Som følger er det nærliggende å tenke at denne reduksjonen i statistisk kraft gjør modellen uegnet, eller iallfall svekket, for å predikere boliger av høyere verdi. Det forsterkes ytterligere av det faktum at nevrale nettverk fordrer en

stor mengde data for tilstrekkelig trening. Dette mulige problemet gjelder imidlertid ikke for modell for førstegangskjøpere, i og med at observasjoner over 5 millioner kroner er fjernet.

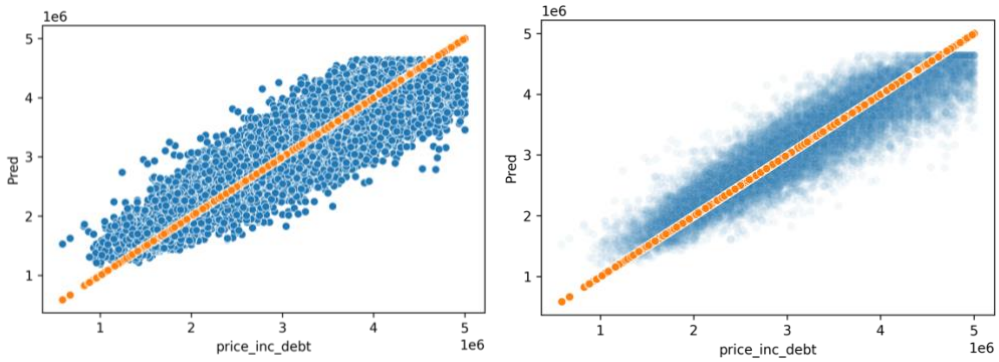


Figur 5: Feilprediksjon ANN Ordinær modell; histogram

I figur 5 ser vi et histogram med fordelingen over feilprediksjon i kroner. Her kan man observere en normalfordeling, men med forventningsverdi på omkring 500 000 kroner. Det vil si at modellen har en tendens til å overestimere.

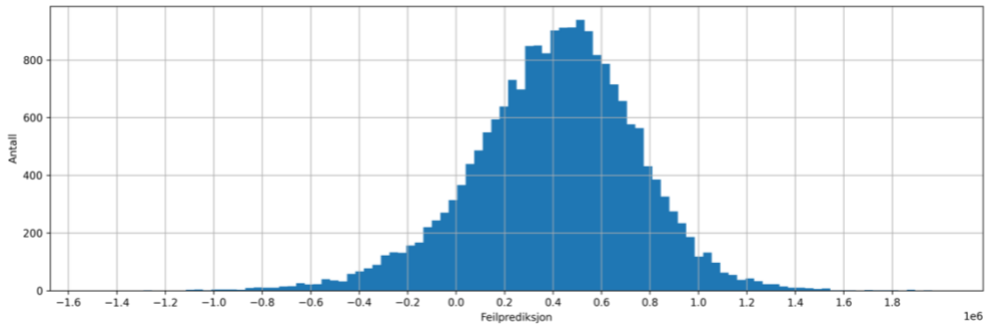
Modell for førstegangskjøper:

Den andre modellen vi lagde, med en makspris på 5 millioner kroner, fikk R^2 på 84,3%. Dette måltallet er imidlertid uegnet for sammenligning med ordinær modell, fordi SS_{TOT} er betydelig redusert som følge av begrensningene vi gjorde på dataene. RMSE for modellen er 343 207 kroner, mens MAE er 266 496 kroner. Disse tallene vil være på en annen skala enn for ordinær modell på grunn av de begrensningene som er gjort. Differansen mellom RMSE og MAE tyder på at det er en viss variasjon i prediksjonsavvikene. MAPE er 9,1%. Det vil si at modell for førstegangskjøpere er mer presis i sine prediksjoner enn den ordinære modellen, som hadde en MAPE på 10,9%.



Figur 6: Feilprediksjon ANN Modell for førstegangskjøpere; spredningsplott

Figur 6 har to spredningsplott med ulik styrke som viser faktisk pris mot predikert pris. I plottet kan man observere homoskedastisitet, i kontrast til den ordinære modellen, der vi observerte noe form av heteroskedastisitet.



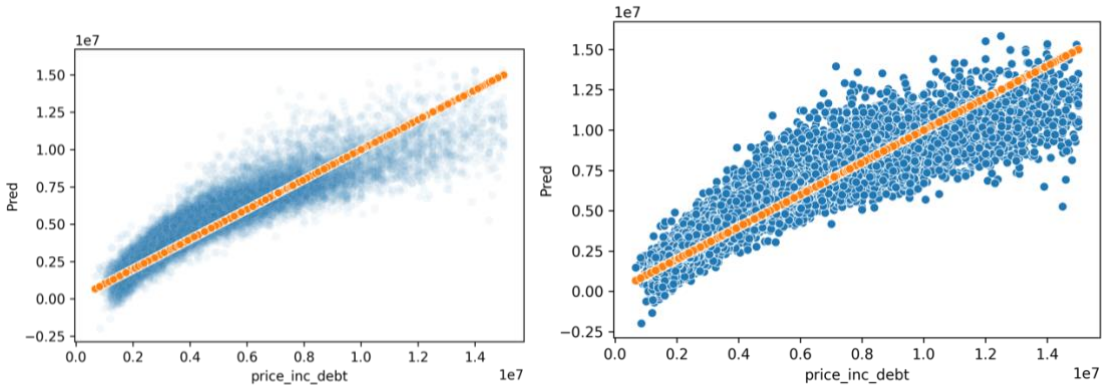
Figur 7: Feilprediksjon ANN Modell for førstegangskjøpere; histogram

Av fordelingen i figur 7 kan man observere en normalfordeling, med forventning omkring 500 000 kroner. Det vil si at modell for førstegangskjøpere på samme måte som ordinær modell har en tendens til å overestimere. En mulig forklaring er at responsvariabelen er nokså høyreskjev.

6.2 Hedonisk prismodell

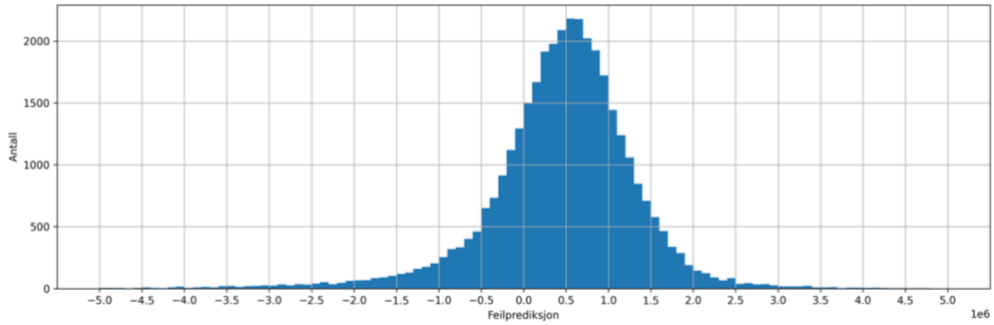
Modell med multippel regresjon:

Modellen har en R^2 på 83,7%. Den ordinære modellen på nevralt nettverk fikk til sammenligning en score på 90,23%. Begge har en makspris på 15 millioner. Det tyder på at det nevrale nettverket i større grad evner å fange opp mønstrene i dataene. RMSE er 909 046 kroner, mens MAE er 628 864 kroner. Altså er det en viss grad av variasjon i avvikene til prediksjonene. MAPE er 15,73%. Det vil si at modellen jevnt over viser dårligere resultater enn for nevrale nettverk.



Figur 8: Feilprediksjon HPM Modell med multippel regresjon; spredningsplott

Av plottet til høyre i figur 8 kan man se at modellen synes å overestimere for boliger av lav verdi, mens den underestimerer for boliger av høy verdi. En mulig forklaring er at den hedoniske modellen er lineær, mens dataene har en ikke-lineær sammenheng med responsvariabelen.



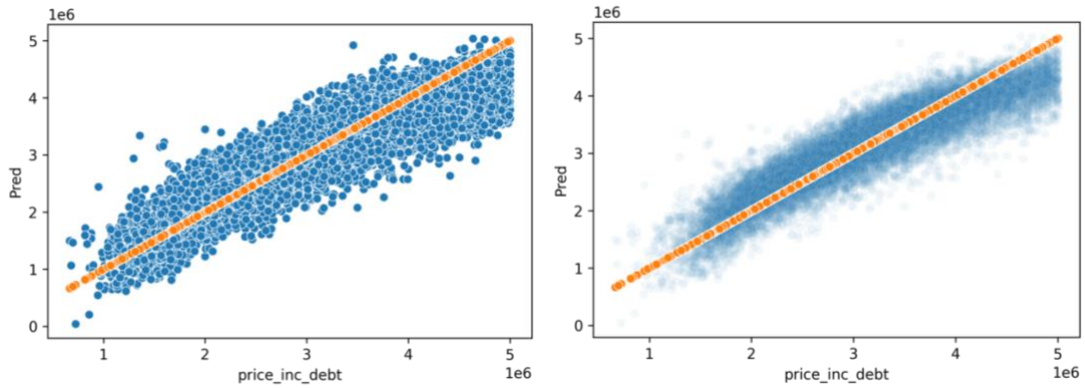
Figur 9: Feilprediksjon HPM Modell med multippel regresjon; histogram

I figur 9 ser vi et histogram over feilprediksjonen for HPM modell med multippel regresjon. Som vi kan observere, er avvikene til prediksjonene normalfordelte med forventning rundt 600 000 kroner.

Modell med logaritmisk transformasjon:

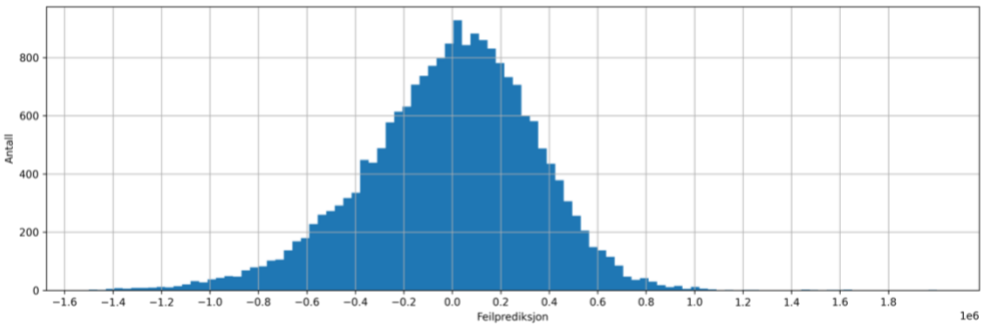
Modellen med logaritmisk transformasjon har samme begrensninger som modell for førstegangskjøpere (ANN), samtidig som den avhengige variabelen har gjennomgått en logaritmisk transformasjon. Modellen har en R^2 på 83,8%. Til sammenligning fikk modellen for førstegangskjøpere under nevralt nettverk en score på 84,3%. RMSE er 349 758 kroner, mens MAE er 272 741 kroner. Modellen for førstegangskjøpere fikk en MAE på 266 496 kroner. Forskjellen er marginal, og det er dårlig grunnlag for å si at det nevrale nettverket har

prestert nevneverdig bedre. MAPE er 9,47%, og dette kan understøtte betraktningen om at det ikke er nevneverdig forskjell i presisjonen til modellen for førstegangskjøpere og modellen med logaritmisk transformasjon.



Figur 10: Feilprediksjon HPM Modell med logaritmisk transformasjon; spredningsplott

I figur 10, slik som i HPM modellen med multippel regresjon modellen, kan vi observere homoskedastisitet.



Figur 11: Feilprediksjon HPM Modell med logaritmisk transformasjon; histogram

Til tross for den logaritmiske transformasjonen i HPM med logaritmisk transformasjon så ble ikke resultatene nevneverdig bedre i forhold til HPM med multippel regresjon. Dette står i kontrast til ordinær modell og modell for førstegangskjøpere, hvor skalering ble brukt på dataen. Kunstig nevrale nettverk er mer avhengig av å ha dataen på samme skala enn det man er ved bruk av regresjon.

6.3 Overordnet diskusjon

Det er på det rene at modellene har ulike styrker og svakheter, og ulike bruksområder. Som følger vil det være forslag til forbedringer for modellene. Denne seksjonen dreier seg om disse styrkene og svakhetene, og mulige forbedringer.

Alle modellene vi har utformet er definert av en avgrensning i dataene. Mens den ordinære modellen og modellen med multippel regresjon har en begrensning på enheter under 15 millioner kroner, har modell for førstegangskjøpere og modell med logaritmisk transformasjon en begrensning på 5 millioner kroner. Av den grunn vil ordinær modell og modell med multippel regresjon være mer fleksible modeller enn de to andre med større begrensninger. På den andre siden er modell for førstegangskjøpere og modell med logaritmisk transformasjon mer presise i sine estimat. Valget av modell vil derfor begrunnes ut ifra hva man anser som viktigst; presisjon eller fleksibilitet.

I oppgaven har vi utforsket hedonisk prismodell og nevrale nettverk. Det er liten tvil om at hedonisk metode er enklere å forstå og benytte seg av. Derfor krever metoden mindre av brukeren av modellen samtidig som at det for visse bruksformål vil være viktig å forstå hvorfor modellen spytter ut estimatene den gir. Likevel har hedonisk prismodell noen svakheter. For det første er det noen markedsrelaterte begrensninger. I hedonisk metode antas det at konsumentene kan tilpasse kombinasjonen av attributter slik de selv ønsker, mens det sjelden fungerer slik i virkeligheten. Modellen fanger ikke opp faktorer som skatt og avgifter, som også spiller inn på om en konsument har mulighet til å kjøpe boligen (Chin & Chau, 2003). Multikollinaritet er også et problem, og det kan tenkes at flere variabler korrelerer med hverandre. For eksempel vil BRA og PROM være avhengige av hverandre, og dette kan forstyrre modellen (Solbakken, 2019, s. 286).

Nevrale nettverk har visse egenskaper som resulterer i styrker og svakheter. For det første er nevrale nettverk en universal approksimeringsmetode. Den hedoniske prismodellen er derimot lineær og dermed uegnet for å fange opp ikke-lineære sammenhenger. Ikke-linearitet er som nevnt en av hovedkritikkene mot den hedoniske modellen. Derfor vil egenskapen som en universell approksimeringsmetode være en styrke for nevrale nettverk, særlig overfor den hedoniske prismodellen. For det andre er nevrale nettverk avhengig av valg av hyperparametere. Det finnes automatiserte metoder som rutenett-søk og tilfeldig søk, imidlertid er dette noe som gjør metoden ytterligere kompleks å sette seg inn i og vil dessuten fordre betydelig maskinkraft. Manuelt søk med prøving og feiling som modellene våres er basert på, synes å være en mindre vitenskapelig metode. Den hedoniske modellen har på sin side ingen hyperparametere, noe som igjen gjør den enklere å ta i bruk.

Datasettet som er benyttet har noen svakheter. Selv om datasettet er stort og omfattende, er det variabler som er utelatt. Særlig makroøkonomiske variabler som styringsrente og inflasjon synes å være aktuelle i så henseende. En forbedring som kunne tenkes å bøte på dette, ville vært å introdusere en prisindeks for hver distrikt i Oslo slik som Birkeland et. al. gjør i sin undersøkelse av automatisk verdsettelse. En leilighet som ble solgt for 2 millioner kroner i 2008, kan i 2022 bli solgt for nærmest det dobbelte. Dette er en prisstigning det er vanskelig å få med i modellen. En annen svakhet med datasettet er at det inneholder flere feilverdier og nullverdier som ikke er håndtert. Sannsynligvis finnes det fortsatt flere feilverdier i datasettet vårt som vi ikke har klart å fange opp. Dette vil påvirke modellen vår negativt.

7.0 Konklusjon

Formålet med oppgaven var å lage en modell ved bruk av maskinlæring for å predikere eiendomspriser i Oslo. Ordinær modell og modell for førstegangskjøpere presterte generelt bedre enn modell med multippel regresjon og modell med logaritmisk transformasjon, hvis vi legger MAPE til grunn. Det tyder på at kunstige nevrale nettverk fungerer bedre til eiendomsprediksjon enn hedonisk regresjonsanalyse. En nærliggende forklaring på dette er at hedonisk metode tar for seg lineære problemer, mens boligprispredikering kan sies å inneholde flere elementer av ikke-linearitet. Altså er nevrale nettverkets egenskap som universal approksimeringsmetode særlig attraktivt.

Det finnes ikke en modell som er perfekt for alle sammenhenger, isteden må man se på hver modells egenskaper og vurdere hvilken som er best egnet til problemet for hånden. Ønsker man en modell hvis virkemåte er enkel å forstå bør man velge en av modellene som benytter hedonisk regresjon. Er presisjon særlig viktig, bør man velge en av modellene med utvidete begrensninger i dataene, og vice versa.

Som diskutert har modellene våre begrensninger og som følger et begrenset gyldighetsområde. Blant annet kan de ikke brukes til næringseiendom, ei heller for dyre og/eller spesielle eiendommer. I tillegg tar ikke modellene for seg makroøkonomiske forhold som kan forstyrre resultatene, som igjen begrenser bruksområdet til modellene. Til videre forskning kan det anbefales å utvikle en modell hvor man tar høyde for prisstigning, og eventuelt andre makroøkonomiske forhold. I tillegg kan en med fordel prøve ut andre typer modeller, det vil si andre maskinlæringsmetoder og nevrale nettverk med andre valg for hyperparametere.

8.0 Referanseliste

Abidoye, R. B & Chan, A. P. C. (2018). *Improving property valuation accuracy: a comparison of hedonic pricing model and artificial neural network.* Tilgjengelig fra https://www.tandfonline.com/doi/pdf/10.1080/14445921.2018.1436306?needAccess=true (Hentet 29.03.2022)

Bae, J. K & Park B. (2015). Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data. Tilgjengelig fra: https://www.sciencedirect.com/science/article/pii/S0957417414007325 (Hentet 29.03.2022)

Barlindhaug, R. (2005). *Storbyens boligmarked.* Tilgjengelig fra: https://www.idunn.no/doi/10.18261/ISSN1504-3045-2005-05-11 (Hentet 10.02.2022)

Becker, D. (2021). Linear Regression. *BBAN3001 Essentials of Business Analytics*. Tilgjengelig fra: https://ntnu.blackboard.com (Hentet 18.03.2022)

Birkeland, K. B., D'Silva, A. D., Füss, R. & Oust, A. The Predictability of House Prices: "Human Against Machine". *International Real Estate* Review 24(2) s. 139-183. Tilgjengelig fra: https://www.gssinst.org/irer/wp-content/uploads/2021/07/v24-no2-1_The-Predictability-of-House-Prices_Human-Against-Machine.pdf. (Hentet 08.04.2022).

Bock, S., Goppold, J., & Weiss, M. (2018). *An improvement of the convergence proof of the ADAM-optimizer* [Conference paper]. OTH-Clusterkonferenz, Weiden, Tyskland. arXiv. Tilgjengelig fra: https://arxiv.org/pdf/1804.10587.pdf (Hentet 25.03.2022)

Buodd, M.F. & Derås, E. J. (2020) *Machine Learning for Property Valuation: An Empirical study of how property price predictions can improve property tax estimations in Norway.* Masteroppgave. NHH Norges Handelshøyskole. Tilgjengelig fra: https://openaccess.nhh.no/nhh-xmlui/bitstream/handle/11250/2739788/masterthesis.pdf?sequence=1&isAllowed=y (Hentet 16.03.2022).

Bydelsfakta – Oslo kommune (2019). Boligpris for blokkleiligheter. Tilgjengelig fra:

https://bydelsfakta.oslo.kommune.no/bydel/alle/boligpriser (Hentet 10.02.2022)

Byrådsavdeling for finans (2021) *Boliger etter bygningstype*. Tilgjengelig fra: https://bydelsfakta.oslo.kommune.no/bydel/alle/bygningstyper (Hentet 07.02.2022)

Chin, T. L. & Chau, K. W. (2003). A critical review of literature on the hedonic price model, *International Journal for Housing and Its Applications*, 27 (2), s. 145-165.

Dvergsdal, H. (2019). *Nevralt Nettverk*. Tilgjengelig fra: https://snl.no/nevralt_nettverk (Hentet 16.02.2022).

Finn. *Finn eiendom*. Tilgjengelig fra: https://www.finn.no/bedriftskunde/eiendom (Hentet 10.02.2022)

Hastie, T., Tibshirani, R. & Friedman, J. (2008). *The Elements of Statistical Learning*. 2. utg. Springer.

Hornik, K, Stinchcombe, M & White, H. (1989). Multilayer Feedforward Neural Networks are Universal Approximators. *Neural Networks*, 2, s. 359-366. Tilgjengelig fra: https://www.sciencedirect.com/science/article/pii/0893608089900208 (Hentet 16.02.2022).

Iversen, J.M.V. (2021). Regresjon – analyse av sammenhenger. *MET2010 Anvendt statistikk*. Tilgjengelig fra: https://ntnu.blackboard.com (Hentet 22.03.2022)

Limsombunchai, V. (2004) House price prediction: Hedonic price model vs. artificial neural network. *New Zealand Agricultural and Resource Economics Society Conference, 25-26 June 2004*. Blenheim, New Zealand: New Zealand Agricultural and Resource Economics Society.

Lokalwiki (2004). *Bydeler i Oslo fra 2004*. Tilgjengelig fra: https://lokalhistoriewiki.no/wiki/Fil:Bydeler-Oslo-2004.png (Hentet 10.02.2022)

Osland, L. (2001) Den hedonistiske metoden og estimering av attributtpriser. *Norsk Økonomisk Tidsskrift*, 115, s. 1 - 22.

Pandit, V og Schuller B, (2019). *The Many-to-Many Mapping Between the Concordance Correlation Coefficient and the Mean Square Error*. Tilgjengelig fra: https://arxiv.org/pdf/1902.05180.pdf (Hentet: 7.04.2022)

Rosen, S. (1974) Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition, *Journal of Political Economics*, 82, s. 34 – 55.

Selim, H. (2009). *Determinants of house prices in Turkey: Hedonic regression versus artificial neural network.* Tilgjengelig fra: https://www.sciencedirect.com/science/article/pii/S0957417408000596?casa_token=JVp5vw

u5b6IAAAAA:DI5ccXd8lX7HCZVuNmuG24NLykmOr4LYI7qZcBvEzzPAR7hlSuHl10tpG2-jpcKgdrNPARlYhw (Hentet 29.03.2022)

Solbakken, S. S. (2019) *Statistikk for nybegynnere*. Bergen: Fagbokforlaget

Tidemann, A & Elster A. C. (2022) Maskinlæring. Store Norske Leksikon. Tilgjengelig fra: https://snl.no/maskinlæring (Hentet 16.03.2022).

Walbækken, M. & Bendictow, A. (2020): *Kvantitativ analyse av boligomsetning og budgivning*. Tilgjengelig fra: https://www.regjeringen.no/contentassets/79704749185444af9cad00cb5f95c17c/samfunnsokono misk-analyse.pdf (Hentet 10.02.2022)

Vedlegg

Vedlegg 1

Variabel-navn	**Beskrivelse**	**Null-verdier**
DocumentDate	Tinglysningsdato	0
Picturedate	Dato bilder i annonsen er lastet opp. Er ukorrekt i tidligere annonser.	257
Lastchanged	Siste gang annonsen ble endret. Noen ganger er dette dagen boligen ble solgt. Andre ganger har annonsen blitt endret i etterkant	0
Adcreated	Estimert dato annonsen ble laget	0
Price	Pris	0
Listprice	Prisantydning	0
Lat	Breddegrad	1568
Lng	Lengdegrad	1568
Address	Adresse	0
Apartmentnumber	Leilighetsnummer – eks: H0101	822
Apartmenttype	Type leilighet, 99% er «B»	792
Conveyancetype	100% er «fritt salg»	0
Finncode	Unik kode for å finne annonse på finn.no	0
BRA	Bruksareal. Det vil si hele arealet på boligen ekskludert ytterveggene.	3723
PROM	Primærrom. Det vil si alle primære rom som soverom, kjøkken, bad, kjellerstue osv. Det er dette arealet man bruker til å beregne kvadratmeterpris.	0
Buildyear	Byggeår	925

Matrikkel	Unikt nummer for hver eiendom. Matrikkelen er Norge sitt offisielle eiendomsregister.	0
Bedrooms	Antall soverom	9385
Rooms	Antall rom	51910
Bathrooms	Antall bad	0
WC	Antall toalett/WC	0
Elevator	Binær variabel kodet med J/N (Ja/Nei)	792
Floor	Hvilken etasje leiligheten ligger i	37977
Housingtype	Type eiendom. 87% er leiligheter	0
Commondebt	Fellesgjeld. Må legges til salgssummen for å finne totalpris.	58689
Commoncosts	Felleskostnader. Månedlige kostnader for borettslag ol.	15759
F_Aircondition	Aircondition i boligen	178173
F_Alarm	Alarm i boligen	175171
F_BalkongTerrasse	Har man balkong/terrasse	60920
F_Barnevennlig	Barnevennlig område?	72380
F_Bredbåndstilknytning	Tilknyttet bredbånd	80143
F_Fellesvaskeri	Fellesvaskeri	148196
F_GarasjeP-plass	Garasje eller parkeringsplass	108973
F_Heis	Heis i boligen	129931
F_Ingen gjenboere	Ingen gjenboere (betyr som oftest at ingen har innsyn i boligen)	132973
F_Lademulighet	Lademulighet for elbil	173472
F_Livsløpsstandard	Livsløpsstandard. Det vil si om boligen er tilrettelagt slik at den kan brukes i alle livets faser, også om man får en bevegelseshemming.	176498
F_Moderne	Moderne	136234

F_Offentlig vannkloakk	Tilkoblet offentlig vann/kloakk	98042
F_Parkett	Parkett i boligen	82952
F_PeisIldsted	Peis/ildsted i boligen	128068
F_Rolig	Rolig område	76928
F_Sentralt	Sentral beliggenhet	81605
F_Utsikt	Utsikt fra boligen	119962
F_Vaktmester-vektertjeneste	Vaktmester og/eller vektertjeneste	117235
F_Bademulighet	Bademulighet i nærheten	172852
F_Fiskemulighet	Fiskemulighet i nærheten	178556
F_Turterreng	Turterreng i nærheten	100291
Adtitle	Annonsetittel	0
N_modernisering	Modernisering	0
N_oppussing	Oppussing	0
N_regulert	Regulert	0
N_garasjemulighet	Garasjemulighet	0
Postcode	Postkode	0
Parking	Parkering ved boligen	0
Price_inc_debt	Pris inkludert fellesgjeld (totalpris)	0
Area_id	Område-ID. Tall mellom 1-16 som representerer en bydel i Oslo. Se liste over bydeler under.	0

**Variabler som starter med F_XXX er dummy-variabler.*

***Variabler som starter med N_XXX er generert fra annonsetittelen og vil være veldig upresise*

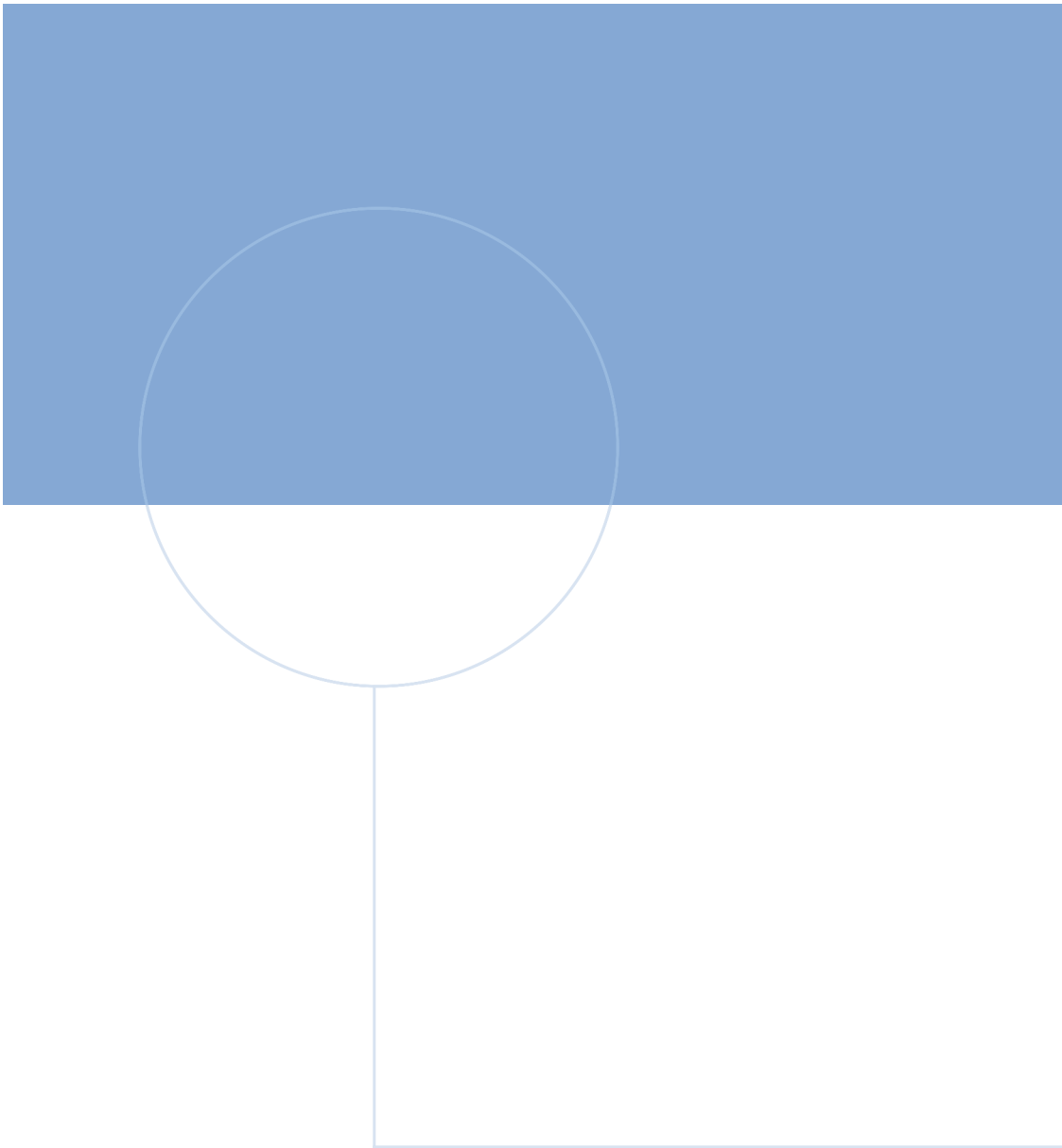
Vedlegg 2

Område-ID	Bydel
1	Gamle Oslo
2	Grünerløkka
3	Sagane
4	St. Hanshaugen
5	Frogner
6	Ullern
7	Vestre Aker
8	Nordre Aker
9	Bjerke
10	Grorud
11	Stovner
12	Alna
13	Østensjø
14	Nordstrand
15	Søndre Nordstrand
16	Oslo sentrum (kvadraturen)

Vedlegg 3

Attributt		Forventet effekt på boligpris
Lokasjon	Avstand fra sentrum	-
	Utsikt til sjø, hav eller elv	+
	Utsikt til fjord/daler/golfbane	+
	Hindret sikt	-
	Lengde på tomteleie	+
Struktur	Antall rom, soverom og baderom	+
	Gulvareal	+
	Kjeller, garasje og veranda	+
	Bygningsservice (f.eks. heis og air condition)	+
	Etasje (bare for leilighetskompleks)	+
	Kvalitet (f.eks. design, materiale og inventar)	+
	Fasiliteter (f.eks. svømmebasseng, treningsrom og tennisbane)	+
	Alder på bygning	-
Nabolag	Inntekten til naboer	+
	Nærhet til gode skoler	+
	Nærhet til sykehus	?
	Nærhet til steder for tilbedelse (f.eks. kirke, moske og templer)	+
	Kriminalitet	-
	Trafikk-/flybråk	-
	Nærhet til kjøpesenter	?
	Nærhet til skog	?
	Miljø (f.eks. landskap, hage og lekeplass)	+

+: positiv innvirkning på boligpris, -: negativ innvirkning på boligpris, ?: varierer fra sted til sted, faktisk effekt er et empirisk spørsmål. Chin & Chau (2003).



NTNU

Kunnskap for en bedre verden