

Ejendomsvurderinger ved brug af XGBoost





AALBORG UNIVERSITET
STUDENTERRAPPORT

Titel:

Ejendomsvurderinger ved brug af XGBoost

Uddannelse:

9. Semester - cand. Oecon.

Social Data Science Semesterprojekt - E2023

Projekt udarbejdet af:

Patrick Nicko Printz

Studienummer: 20195998

Vejleder:

Roman Jurowetzki

Anslag inkl. mellemrum: 78.798

Antal normalsider: 32,8 sider

Afleveringsdato: onsdag d. 20. december 2023

Læsevejledning

Projektets resultater er baseret på kodning i Python 3.11.5, hvoraf koden er uploadet på GitHub og kan findes på linket: <https://github.com/PatrickPrintz/Ejendomsvurdering-XGBoost/tree/main>. På GitHub findes ligeledes link til det anvendte datasæt, samt trænede modeller, der er uploadet på Google Drev. Der er tillige udviklet en illustrativ app, der for rækkehus beskriver hvordan enkelte vurderinger kan forklares. Linket til denne fremgår ligeledes af ovenstående link.

Indholdsfortegnelse

1	Introduktion	1
2	Problemformulering	2
3	Afgrænsning	2
4	Baggrund og intentionerne bag nye ejendomsvurderinger.....	3
4.1	<i>Baggrunden for de nye ejendomsvurderinger.....</i>	3
4.2	<i>Grundprincipperne bag det nye ejendomsvurderingssystem</i>	5
5	Data og metode.....	9
5.1	<i>Data.....</i>	9
5.2	<i>Dataforberedelse, feature-engineering og beregninger.....</i>	10
5.3	<i>Deskriptiv beskrivelse af dataet</i>	12
5.4	<i>Metode.....</i>	18
5.5	<i>Afvejning mellem træfsikkerhed og gennemsigtighed.....</i>	21
5.6	<i>Modelopbygning.....</i>	23
6	Resultater	24
6.1	<i>Evaluering og sammenligning med Engbergudvalgets modeller.....</i>	25
6.2	<i>SHAP-værdier og gennemsigtighed</i>	32
6.3	<i>Opsummering af resultater</i>	37
7	Diskussion	39
7.1	<i>Kan projektets og Engbergudvalgets modeller sammenlignes?</i>	39
7.2	<i>Kan den fornødne gennemsigtighed opretholdes, hvis man benytter XGBoost?</i>	41
8	Konklusion	44
9	Bibliografi	46
10	Appendiks	49
10.1	<i>Dataindsamlingsprocessen</i>	49
10.2	<i>Beskrivelse af variable i datagrundlaget</i>	51
10.3	<i>Logaritmisk transformation af handelspriserne</i>	54
10.4	<i>Regneeksempel på to ejendomme ved brug af Engbergudvalgets model.....</i>	55

Oversigt over figurer

Figur 1 - Kort over placering af enfamiliehuse, ejerlejligheder og rækkehuse i datasæt	14
Figur 2 - Fordeling af handelspriser og logaritmisk transformation	15
Figur 3 - Sammenligning mellem fordelingen af afvigelser for enfamiliehuse	27
Figur 4 - Sammenligning mellem fordelingen af afvigelser for rækkehuse	29
Figur 5 - Sammenligning mellem fordelingen af afvigelser for ejerlejligheder	31
Figur 6 - Variable rangeret efter indflydelse på vurdering	33
Figur 7 - Kort over træfsikkerhed for alle modeller på kommunalt niveau	34
Figur 8 - Numeriske variable og SHAP-værdi	35
Figur 9 - SHAP-værdier for antal badeværelser og toiletter efter energimærke	37

Oversigt over tabeller

Tabel 1 - Den gamle ejendomsvurderingssystems udfordringer	4
Tabel 2 - Engbergudvalgets modeltyper	6
Tabel 3 - Variabler i datagrundlaget og type	10
Tabel 4 - Boligtyper og antal	12
Tabel 5 - Fordeling af boligtype efter region	13
Tabel 6 - Overblik over distancemål	16
Tabel 7 - Fordeling af udvalgte kategoriske variable	17
Tabel 8 - Eksempel på beslutningstræ	19
Tabel 9 - Træning- og testsæt størrelse	24
Tabel 10 - Resultat af hyperparameter tuning	24
Tabel 11 - Sammenligning af modeller for enfamiliehuse	25
Tabel 12 - Over- og underestimering for enfamiliehuse	26
Tabel 13 - Sammenligning af modeller for rækkehuse	27
Tabel 14 - Over og underestimering for rækkehuse	28
Tabel 15 - Sammenligning af modeller for ejerlejligheder	30
Tabel 16 - Over og underestimering for ejerlejligheder	30
Tabel 17 - Beregningseksempel på to sammenlignelige boliger baseret på SHAP-værdier	36
Tabel 18 - Sammenligning af testsæt	39

Oversigt over appendiks

Appendiks 11-1 - Dataindsamling grafik	50
Appendiks 11-2 - Databeskrivelse	51
Appendiks 11-3 - Sammenligning af evalueringsparametre ved logaritmisk transformation	54
Appendiks 11-4 - Beregningseksempel på to delvist ens ejendomme ved Engbergudvalgets model	55

1 Introduktion

Boligpriser og -markedet spiller en stor rolle for både boligejers privatøkonomi og den samlede makroøkonomi i Danmark. Derfor forekommer det heller ikke besynderligt, at SKATs ejendomsvurderinger har været genstand for stor omtale og kritik gennem de sidste ti til 15 år. Ejendomsvurderingerne danner grundlaget for beskatningen af fast ejendom, og det er derfor yderst essentielt, at vurderingerne stemmer tilstrækkeligt overens med den faktiske værdi. I 2013 blev det gamle vurderingssystem suspenderet, blandt andet som følge af en flersidet kritik af systemets træfsikkerhed og gennemsigtighed. I forbindelse med suspenderingen blev der nedsat et ekspertudvalg, der havde til formål at undersøge, hvorvidt det var muligt at skabe et nyt, mere træfsikkert og transparent system. Ekspertudvalget fremlagde en model med højere træfsikkerhed, baseret på forklarlige objektive data. Denne model danner grundlag for det nye ejendomsvurderingssystem, og dermed de foreløbige vurderinger af ejerboliger i 2023. På baggrund af disse vurderinger er systemet dog blevet kritiseret for at have en utilstrækkelig træfsikkerhed, og det synes ligeledes uigennemskueligt, hvordan de enkelte vurderinger er udmønstret. Kritikken omhandler blandt andet, at omrent hver tredje ejendomsvurdering afviger med mere end ± 20 pct. fra den faktiske handelspris, og at det er umuligt for den enkelte boligejer at forstå, hvordan vurderingerne er udarbejdet.

At forudsige en ejendoms værdi er en besværlig opgave. Det kræver en stor mængde forskellige data, herunder data om boligens stand og størrelse, områdets karakteristika samt afstanden til forskelligartede interesseområder, mens indflydelsen fra disse sandsynligvis ikke ens for alle ejendomme på tværs af landet. Sammenhængen mellem alle relevante faktorer og en ejendoms handelspris er således ikke en lineær, men kompleks opgave. Som følge heraf er det essentielt, at det anvendte modelværktøj formår at opfange kompleksiteten af ejendommens værdi. Med afsæt i den fornævnte lave træfsikkerhed, kan der derfor sættes spørgsmålstege ved, om den anvendte model til de nye ejendomsvurderingerne er i stand til dette, eller om træfsikkerheden vil kunne øges som følge af komplekse maskinlæringsmodeller? I takt med at kritikken tilmed har omhandlet gennemsigtigheden af de individuelle vurderinger, er det ligeledes væsentligt, at den øgede kompleksitet ikke forringer gennemsigtigheden. Når Skattemyndigheden indkræver skatter på baggrund af vurderingerne, skal disse kunne forklare og samtidig stemme overens med den faktiske værdi. I forlængelse heraf kan der sættes spørgsmålstege ved, om en uigennemskuelig vurdering, der blot rammer indenfor ± 20 pct. i

to ud af tre tilfælde er tilfredsstillende, eller om vurderingerne er bedst forklaret ved mere komplekse metoder?

2 Problemformulering

Med afsæt i ovenstående introduktion vil nærværende projekt søge at besvare nedenstående problemstilling, der lyder:

Er det muligt at forbedre træfsikkerheden af det nye ejendomsvurderingssystem for enfamiliehuse, ejerlejligheder og rækkehuse i Danmark ved brug af maskinlærings algoritmen XGBoost?

Der er en tendens til at opfatte maskinlæringsmodeller som sort-boks-metoder, hvor resultaterne kan være besværlige at forklare. Da nærværende projekt forsøger at udvikle et offentligt værktøj, vil følgende problemstilling ligeledes undersøges:

Er det muligt at forbedre træfsikkerheden, uden at gå på kompromis med gennemsigtigheden?

Projektet har således en todelt målsætning om både at øge træfsikkerheden og gennemsigtigheden af det nye ejendomsvurderingssystem. I forbindelse med dette har det været nødvendigt at tage nogle forbehold, hvilket vil fremgå af næstkommande afsnit.

3 Afgrænsning

Nærværende projekt forsøger at forbedre et relativt nyt system, der endnu ikke er fuldt ud implementeret og dokumenteret på tidspunktet af projektets udarbejdelse. Der er blot fremlagt resultatet af de foreløbige vurderinger, der udgør beskatningsgrundlaget for ejerboliger i 2024. Dokumentationen og den fulde sammensætning af disse er ikke offentliggjort eller mulig at afprøve. I nærværende projekt tages derfor udgangspunkt i resultaterne af den ekspertgruppe, der blev nedsat i 2013, kaldt *Engbergudvalget*. Dette er relevant, da Skatteministeriet (2016) påpeger, at det nye ejendomsvurderingssystem kommer til at ligne dette i stort omfang (Skatteministeriet, 2016, s. 28). Projektet tager forbehold for, at der i fremtiden kan være væsentlige ændringer mellem ekspertgruppens og det nye ejendomsvurderingssystem, som medfører, at projektets resultater ikke er sammenlignelige med det nye ejendomssystem.

4 Baggrund og intentionerne bag nye ejendomsvurderinger

Formålet med dette afsnit er at introducere baggrunden for udviklingen af det nye boligsystem ved at fremhæve kritikken og problematikkerne ved det gamle vurderingssystem. Endvidere introduceres grundprincipperne og tankegangen bag udviklingen af det nye vurderingssystem, samt den fornyede kritik heraf.

4.1 Baggrunden for de nye ejendomsvurderinger

Den fundamentale årsag til revideringen af det gamle ejendomsvurderingssystem kom efter en flersidet kritik af specielt træfsikkerheden af davarende vurderinger. Ejendomsvurderingerne danner grundlaget for beskatningen af boliger i Danmark, hvorfor både høje og lave vurderinger er med til at skævvride beskatningsgrundlaget for den enkelte bolig. Dette kan i flere tilfælde få konsekvenser for den enkeltes privatøkonomi såvel som den samlede skatteindtægt.

Jf. Vurderingslovens §6 vurderes ejendomme på baggrund af værdien i handel og vandel (markedsværdien). Vurderingerne af ejerboliger skal ligge under salgsprisen, da overskridelse heraf vil medføre et uretfærdigt beskatningsgrundlag. Rigsrevisionen fremlagde i 2013, på eget initiativ, en samlet kritik af ejendomsvurderingerne, der påpegede, at 41 pct. af parcelhuse i andet halvår af 2011 blev vurderet over salgsprisen. Endvidere viste deres analyse, at 34 pct. af parcelhusene var undervurderet, selv efter at have frataget 15 pct. af salgsprisen som følge af den accepteret fejlmargin (Rigsrevisionen, 2013, s. 20).

Kritikken fra Rigsrevisionen medførte, at ejendomsvurderingerne blev suspenderet i 2013, og det har derfor været 2011-vurderingerne, der har gjort sig gældende for ejerboliger siden hen. De offentlige vurderinger af ejerboliger blev frem til 2011 fortaget på baggrund af en blanding af analyser af salgspriser samt den enkelte medarbejders faglige vurdering og erfaring. Det subjektive element af vurderingerne resulterede i, at det ikke var muligt for andre at reproducere resultaterne af vurderingerne. De statistiske modeller, der blev anvendt til vurderingerne, blev derfor omtalt som forslagsmodeller, da de blot udgjorde et grundlag for yderligere vurderingsfaglig behandling (DØRS, 2016, s. 250-255).

Principielt består ejendomsvurderingerne af en grund- og bygningsværdifastsættelse. Grundværdierne skal vurderes ud fra antagelsen om, at grunden er ubebygget, hvilket har vist sig besværligt, da flere områder allerede er udbygget og mængden af ubebyggede grunde er mangelfuld

(DØRS, 2016, s. 250). Når grundværdien er fastlagt, skal bygningsværdierne fastlægges. Til dette tages der udgangspunkt i forskelsværdien mellem hele ejendommen og grundværdien, hvortil de seneste fire års salgspriser anvendes som grundlag. I denne proces indgår regressionsmodeller, der tillige benyttes som forslagsmodeller, indeholdende karakteristika ved den enkelte bygning, herunder opførselsår, størrelse, varmeinstallation, ydermur- og tagmateriale. Disse vurderinger er yderligere genstand for efterfølgende vurderingsfaglig behandling. Overordnet fremgår vurderingsprocessen for ejerboliger som en blanding af objektive registerdata og regressionsmodeller, samt manuel vurderingsfaglig bearbejdning baseret på subjektive skøn. Det vurderingsfaglige element resulterer i, at det er umuligt at genskabe og retfærdiggøre, hvilke objektive og subjektive elementer, der har bidraget til den enkelte ejendomsvurdering (DØRS, 2016, s. 252 og Skatteministeriet, 2021, s. 163).

Efter Rigsrevisionens kritik blev der istandsat et ekspertudvalg, Engbergudvalget, i 2014. Udvalget havde til formål at udvikle et vurderingssystem, der i højere grad fokuserer på objektive data og statistisk modellering. Engbergudvalgets arbejde resulterede i en prototype af et system, som har højere træfsikkerhed og er baseret på objektive data. I forlængelse af Engbergudvalget blev et Implementeringscenter for ejendomsvurderinger (ICE) etableret under Skatteministeriet med henblik på at etablere og implementere det nye ejendomsvurderingssystem (Skatteministeriet, 2016, s. 1-3).

ICE og Engbergudvalget identificerede nogle udfordringer ved det gamle ejendomsvurderingssystem, hvilke opsummeres i nedenstående tabel 1.

Tabel 1 - Den gamle ejendomsvurderingssystems udfordringer

Lav træfsikkerhed	Jf. Rigsrevisionens og Engbergudvalgets kritik af nøjagtigheden på vurderinger af ejerboliger.
Uensartede vurderinger	Engbergudvalget fandt flere eksempler på identiske boliger, med forskellige vurderinger.
Gennemsigtigheden mangler	I forlængelse af den manglende ensartethed, er boligvurderingerne besværlige at forklare for den enkelte boligejer.
Skævvridninger i grundvurderinger	Det gamle vurderingssystem resulterede i skævheder mellem grundvurderinger for enfamiliehuse og ejerlejligheder. Dette betyder oftest, at ejere af lejligheder ender med at betale en relativt lavere grundskyld.

For lave grundvurderinger

Det gamle vurderingssystem giver i store områder af landet for lave vurderinger af grunde.

Kilde: Baseret på Skatteministeriet (2016 & 2014)

Udfordringerne ved det gamle vurderingssystem kan i al sin enkelhed karakteriseres ved lav træfsikkerhed og manglende gennemsigtighed, der samlet kan kobles til den subjektive vurdering, der indgår i den endelige vurdering. Træfsikkerhed såvel som gennemskuelighed er derfor også Skatteministeriets pejlemærker for udarbejdelsen af det nye ejendomsvurderingssystem.

4.2 Grundprincipperne bag det nye ejendomsvurderingssystem

På tidspunktet af nærværende projekt, findes der ingen officiel eller udførlig dokumentation fra ICE, Skatteministeriet eller Vurderingsstyrelsen vedrørende specifikke detaljer af modelsammensætningen eller kildekoden brugt til de foreløbige (eller endelige) vurderinger. Af Skatteministeriets udlæg om ”Nye og mere retvisende ejendomsvurderinger” fra 2016 fremgår det, at modellen er baseret på samme metodologiske overvejelser, der blev fremlagt i Engbergudvalgets prototype. En lang række væsentlige ændringer adskiller dog denne model fra både Engbergudvalgets og det gamle ejendomsvurderingssystem (Skatteministeriet, 2016, s. 28). Af hensyn til dette tager nærværende projekt udgangspunkt i Skatteministeriets nuværende og delvist upræcise forklaring af de nye vurderinger, samt Engbergudvalgets prototype, som i hovedtræk må forventes at ligne den endelige model.

4.2.1 Engbergudvalgets prototype¹

Hovedpræmissen for udvalgets forslag til en vurderingsmodel er, at de skal baseres på objektive data, der kan anvendes i en statistisk sammenhæng uden behov for vurderingsfaglige skøn. Foruden præsentationen af en alternativ fremgangsmetode, anbefaler udvalget tillige flere forbedringer af ejendomsvurderings-processen, herunder en nødvendig forbedret datakvalitet, bedre klagemuligheder og større gennemsigtighed for den almene boligejer.

¹ I delafsnittet vil Engbergudvalget være omtalt som udvalget.

Der fremføres fem bud på alternative modeller, der alle adskiller sig fra det gamle vurderingssystem ved at inkluderer geografiske data og naboområdets pris. De fem præsenterede modeller er i stort omfang baseret på Generalized additive modeller (GAM) og er præsenteret i nedenstående tabel;

Tabel 2 - Engbergudvalgets modeltyper

Model	Beskrivelse
En GAM uden nabopriser	En regression af de faktiske BBR-oplysninger om hver enkelt ejendom samt geografiske data. Denne model beregner, hvor meget hver enkelt karakteristika tilføjer af værdi til den enkelte bolig (eksempelvis den ekstra værdi af 1 kvadratmeter boligareal) og er generel for alle ejendomme.
Nabopriser alene	En simpel model, der alene beregner værdien af en bolig ud fra produktet af nabopriserne kvadratmeterpris og ejendommens areal.
En GAM med nabopriser (Anbefalede model)	En variant af første model, som indeholder BBR-oplysninger, geografiske data og nabopriserne. Nabopriserne fremskrives til nutidige priser og indgår sammen med BBR- og geografisk data.
En GAM med nabopriser og historiske priser	En variant af tredje modeltype, der yderligere tilføjer tidlige handelspriser på ejendommen inden for en tidsramme på 6 år. Er ejendommen ikke solgt inden for seneste 6 år, bruges nabopriserne i stedet.
En Generalized Multiplicative Model med nabopriser	Endnu en variant af tredje model, der i stedet for en additiv regressionmodel benytter en multiplikativ regressionstilgang. Den beregnede gennemsnitspris ad nabopriserne justeres ved at gange værdier ejendomsspecifikke karakteristika.

Kilde: (Skatteministeriet, 2014, s. 105)

På baggrund af tre statistiske vurderingsparametre anbefaler udvalget, at der arbejdes videre med GAM-modellen med nabopriser, da denne har den højeste træfsikkerhed. Prototypen af denne model er baseret på historiske handler i perioden år 2001 til 2013. Estimeringen forløber i tre trin; først fremskrives de historiske salgspriser, dernæst estimeres nabopriserne kvadratmeterpris, hvor den endelige pris slutteligt estimeres ved at justere for ejendomsspecifikke karakteristika. Fremskrivningen beror på lokale prisudviklinger, hvilket betyder, at en ejendom solgt i 2008 fremskrives til 2010-priser ved at se på de 50 nærmeste salg i 2008 og 2010 og justere prisen med forskellen i mediankvadratmeterprisen. Nabopriserne beregnes ved mediankvadratmeterprisen af de 18 nærmeste salgspriser for parcelhuse. For ejerlejligheder og rækkehuse benyttes de ti nærmeste

salgspriser. Herefter ganges prisen med den specifikke ejendoms boligareal. Med udgangspunkt i produktet af boligens areal og nabopriserne, korrigeres ejendommenes vurdering efter ejendomsspecifikke karakteristika. Denne regression baseres på 38 variable for parcelhusmodellen, otte for rækkehuse og 11 for ejerlejligheder. Værdierne af disse fratrækkes eller tillægges naboprisestimatet og er generelle på tværs af alle ejendomme af samme type (Skatteministeriet, 2014, s.106-118).

4.2.2 Overgangen til det nye vurderingssystem og kritikken heraf

I Skatteministeriet (2016, s. 28) fremgår det, at det nye vurderingssystem er baseret på Engbergudvalgets anbefalede model, men ”der er dog foretaget en lang række væsentlige ændringer af Engbergudvalgets prototype, ligesom modellen også adskiller sig fra SKATs nuværende model”² (Skatteministeriet, 2016, s. 28, 9-10). Af Skatteministeriets (2016) udlæg fremgår det eksempelvis, at beregningerne af kvadratmeterprisen involverer eventuelle tidlige salg af den specifikke ejendom. Hvordan modellen yderligere adskiller sig fra Engbergudvalgets, er dog ikke præciseret.

De foreløbige (og nye) ejendomsvurderinger bygger således også på den fremskudte kvadratmeterpris i naboområdet, som korrigeres efter den specifikke ejendoms karakteristika og tidlige salg, der til sidst skaleres op til den specifikke ejendoms størrelse (Vurderingsstyrelsen, 2023). På trods af at denne metode fremstår intuitiv og ligeledes ligner Engbergudvalgets anbefalede model, har resultatet af de foreløbige vurderinger medvirket til en flersidet kritik af både træfsikkerheden og gennemskueligheden.

I tiden efter udsendelsen af de foreløbige vurderinger begyndte flere medier at fremfører historier om systemets lave træfsikkerhed baseret på afvigelser fra nylige salgspriser og uigenemskuelige vurderinger. En gennemgang foretaget af DR viser eksempelvis at mere end hver tredje ejendom er fejlvurderet (Ingvorsen, Kielgast, Hecklen, & Ussing, 2023). DR har gennemgået 82 tusinde foreløbige vurderinger af ejendomme, der er solgt inden for seks mdr. af den dato vurderingerne sigter efter. Undersøgelsen påpeger at mere end hver tredje vurdering af disse mere end ±20 pct. ved siden af den pris ejendommen blev solgt for seks måneder før eller efter. ±20 pct. målet er ligeledes et af de paramenter, som Engbergudvalget benytter og illustrerer derfor, hvor lav træfsikkerhed det nye

² Af ’nuværende model’ refereres der, i projektets terminologi, til det gamle ejendomsvurderingssystem, eftersom rapporten kom i oktober 2016, dvs. før implementeringen af nyt system.

vurderingssystem har (Skatteministeriet, 2014, s. 107). Af Skatteministeriets (2016) udlæg fremgår det, at det nye vurderingssystem er etableret på baggrund af dets forgængeres mangler, hvilket indebærer en højere træfsikkerhed og større gennemskuelighed. Blandt boligejerne af de 82 tusinde ejendomme, er det dog ikke synligt, hvordan eller hvorfor deres ejendom vurderes til langt over den nylige handelspris. Dette forekommer specielt i 80 tusinde af vurderingerne, hvor grundværdien er større end den samlede ejendoms værdi, hvilket vil sige at ejendommen er mere værd uden boligen (Bech-Nielsen, 2023)³. Dette stemmer ikke umiddelbart overens med Skatteministeriets (2016) ønske om at højne gennemsigtigheden for den enkelte boligejer. Dette er ligeledes en pointe, der kritiseres af professor i Statskundskab Micheal Petersen, der mener, at beregningerne til grund for disse skæve vurderinger bør offentliggøres (Bech-Nielsen, 2023). Petersen mener, at når systemet har store økonomiske konsekvenser og samtidig kritiseres i så stort omfang, bør gennemsigtigheden fra det offentlige være bedre. Petersen påpeger tillige, at der mangler gennemsigtighed for den enkelte boligejer, men ligeledes også gennemsigtighed i form af en akademisk publikation, som gør det muligt for eksperter at gå systemet i sømmene (Bech-Nielsen, 2023).

4.2.3 Opsummering

På baggrund af ovenstående gennemgang kan det udledes, at det gamle vurderingssystem blev suspenderet, som følge af en flersidet kritik af træfsikkerheden og gennemsigtigheden. Engbergudvalget, der blev nedsat som følge heraf, illustrerede, hvordan det var muligt at forøge begge målsætninger ved brug af objektive data. Det nye vurderingssystem, der i sort omfang minder om udvalgets anbefalinger, har dog ikke umiddelbart forhøjet træfsikkerheden eller gennemsigtigheden, baseret på den umiddelbare kritik af de foreløbige vurderinger.

³ Boligværdi + grundværdi = ejendoms værdi

5 Data og metode

Følgende afsnit har til formål at beskrive det anvendte datagrundlag og indsamlingen heraf, efterfulgt af en introduktion til de anvendte metoder og modeller. Metodeafsnittet vil forholde sig til de praktiske implikationer af den valgte modeltype.

Det salgsdata, som benyttes i projektet, er på egen hånd indhentet fra Boliga.dk. Der er efterfølgende udviklet automatiske løsninger, der kan kombinere dette data med offentlige registerdata. Denne metode har resulteret i, at indsamlingen har været en udfordrende og kompliceret proces, der var nødvendig for at besvare den overordnede problemstilling. Da dataindsamlingen spiller en fundamental rolle i udformningen af nærværende projekt, er beskrivelser af konstruktionen af datagrundlaget inkluderet i et særskilt afsnit i projektets appendiks 11-1. Nedenstående afsnit vil derfor forholde sig overordnet til dataindsamlingen. Endvidere vil der i metodeafsnittet lægges vægt på at beskrive den praktiske anvendelse af metoden, fremfor den dybere matematiske formulering heraf. Ydermere har projektet til formål at opstille én model for hver boligtype i projektets problemstilling. Derfor vil datagennemgangen have en målsætning om at illustrerer forskelligheder mellem disse.

5.1 Data

For at kunne forudsige ejendommes faktiske handelspriser baseret på objektive data, er kvaliteten af det anvendte data særligt vigtig. Derfor har dataindsamlingen i nærværende projekt haft til formål at fremskaffe så mange relevante variable, der kan være med til at forudsige de faktiske markedspriser. Der tages således afsæt i realiserede handelspriser i frit marked baseret på solgte ejendomme i perioden januar 2000 til november 2023, med en pris på over 100.000 kr., ved alm. frit salg. Oplysningerne stammer fra Boliga.dk og udgør indledningsvist i alt 1.144.150 enfamiliehuse, ejerlejligheder, rækkehuse og landejendomme. Da landejendomme ikke er i projektets interesse frasorteres disse.

Salgsdataet er efterfølgende suppleret med en lang række af offentlige registeroplysninger omkring grund- og boligkarakteristika, der beskriver de mest relevante forhold, såsom grundstørrelse, energimærkning, varmekilde m.m. Modsat Engbergudvalget er der i nærværende projekt inkluderet 30 års-renten af realkreditlån. Årsagen til dette skal findes ved den betydelige rolle, som denne har for den generelle prisudvikling på boligmarkedet (Nationalbanken, 2021, s. 4).

Som led af projektets *feature-engineering* beregnes en lang række distance mål baseret på geografiske forhold, samt vægtede gennemsnitspriser af nærmeste ejendomme. Det forventes, at geografiske forhold spiller en mærkbar rolle, og at de vægtede gennemsnitspriser ligeledes giver en klar indikation på den specifikke ejendoms mulige handelspris. Geografiske forhold og nabopriser er ligeledes inkluderet i Engbergudvalgets prototype, der tillige benytter afstande til forskellige interesseområder.

Det endelige datagrundlag består således af en lang række oplysninger omkring bolig- og grundkarakteristika, geografiske forhold og distancer til interessepunkter samt en enkelt makroøkonomisk variabel. Nedenstående tabel 3 beskriver det data, som anvendes i projektets modeller. I appendiks 11-1 er dataindsamlingen beskrevet mere udførligt, herunder hvordan dataet er fremskaffet, og appendiks 11-2 beskriver hvad variablene indebærer.

Tabel 3 - Variabler i datagrundlaget og type

Variable	Type	Variable	Type
Handelspris (Target)	Numerisk	Boligens anvendelse	Kategorisk
Antal værelser	Numerisk	Energimærkning	Kategorisk
Opførstselsår	Numerisk	Grundstørrelse	Numerisk
Boligareal	Numerisk	Sogn	Kategorisk
Region	Kategorisk	Varmeinstallation	Kategorisk
Zone	Kategorisk	Ydervæg materiale	Kategorisk
Tagtype	Kategorisk	30 års rente	Numerisk
Distance til motorvej	Numerisk	Gennemsnitspris for 20 nærmeste ejendomme	Numerisk
Distance til lufthavn	Numerisk	Distance til vandløb, sø m.m.	Numerisk
Distance til kyst	Numerisk	Distance til skov	Numerisk
Distance til togspor	Numerisk	Distance til børnehave og daginstitutioner	Numerisk
Distance til universitet / forsknings institut	Numerisk	Distance til ungdomsuddannelse og folkeskole	Numerisk
Længdegrad	Numerisk	Breddegrad	Numerisk

Annotering: Der refereres til projektets databeskrivelse i appendiks 1 for uddybning af hvad variable indebærer.

5.2 Dataforberedelse, feature-engineering og beregninger

Eftersom handelspriserne i dataet er over tid, benyttes forbrugerprisindeks til at fremskrive priserne til 2022-priser. Dette udføres ved brug af data fra Danmarks Statistik for forbrugerpriserne opgjort årligt. For 2023 benyttes gennemsnittet til og med oktober. Ydermere udføres en filtrering af dataet med afsæt i Adolfsen, et al. (2022, s. 20), som anvender en lignende dataindsamling og som opstiller kriterier, der indsnævrer dataet til at indeholde relevante boliger. Filtreringen indebærer at ejendommen er solgt for en pris under 100 mio.kr., har boligareal under 1000 m^2 , er bygget efter år

1200, har mindre antal værelser end 20 og mindre end 10 badeværelser og toiletter. Filtreringen suppleres ved at fjerne store afvigelser ved brug Z-scoren for de fremskrevne priser, hvor handelspriser med en Z-score over tre fjernes. De fjernede ejendomme vil sandsynligvis være ejendomme, hvis vurdering burde suppleres med vurderingsfaglig behandling (Skatteministeriet, 2014, s. 9).

Endvidere fjernes observationer, der har manglende registeroplysninger, såsom ingen tag- eller vægmaterialekode, samt boliger uden en opgjort energimærkning. Manglende energimærkning reducerer alene antallet af observationer med 298.909 ejendomme, og det endelige data udgør derfor et samlet antal observationer på 655.563.

5.2.1 Feature-engineering

Dataet vedrørende geografiske data, såsom distancer, er beregnes ved brug af kortlægning over placeringen af lufthavne, togspor, motorveje, Danmarks kyst, skoler og institutioner, vandløb samt skovområder. Dertil beregnes distancen mellem hver bolig, baseret på længde- og breddegrader, og nærmeste punkt for de fornævnte steder. For områder som skov og vandløb beregnes distancen mellem boligen og midtpunktet for område-polygonen. Distancerne er opgjort i kilometer og beregnes ved *Haversine-formlen* målt i fugleflugtslinje ved brug af koordinatsystemet World Geodetic System 1984 (WGS84). Det er vigtigt at påpege, at distancerne er approksimerede afstande og kan forbedres, men det antages at eventuel forbedring heraf er ubetydelig (Kettle, 2017).

For gennemsnitsprisen af de 20 geografisk nærmeste ejendomme, vægtes priserne efter afstand til den specifikke ejendom. De eneste kriterier for dette er, at den specifikke ejendom og referenceejendomme er i samme region. Dertil beregnes vægtningen af hver referenceejendom, ved formlen:

$$w_i = \frac{1}{Distance\ i\ km} + (1e - 6)$$

Hvortil gennemsnitsprisen beregnes ved:

$$Gennemsnitspris_{20} = \frac{\sum_{i=1}^{20} (w_i * P_i)}{\sum_{i=1}^{20} w_i}$$

Vægtningen bevirket, at ejendomme geografisk tættest på får højst vægtning og ejendomme længst væk får lavest vægtning. Vægtene tilføjes $1e - 6$ for at sikre at ejendomme med distancer på nul kilometer kan beregnes og får en høj vægtning. Dette vil typisk være ejerlejligheder i samme opgang.

I projektets indledende fase var det valgt at udvælge referenceejendomme baseret på flere kriterier end samme region. Disse kriterier indebærer, at den specifikke ejendom og reference har/er; (1) I samme sogn og region, (2) samme varmeinstallation, (3) samme antal toiletter og badeværelser ± 2 , (4) samme boligtype, (5) samme antal værelser ± 2 , (6) samme grund- og boligareal ± 30 pct., (7) solgt indenfor seks år og (8) samme energimærkning. Disse kriterier resulterer dog i, at flere ejendomme har få til ingen referenceejendomme. Restriktionerne skulle lempes betydeligt, før dette er muligt at udføre og sikre, at hver ejendom har minimum fire referenceejendomme. Derfor er det fra projektets side valgt, at priserne vægtes efter distancerne i stedet.

Overordnet synes dataet at være rig på flere relevante faktorer og ligeledes være et godt udgangspunkt for et lignende datagrundlag, som Engbergudvalget benytter. Ydermere synes dataet at være af høj kvalitet, hvilket er nødvendigt for at opnå en træfsikker model. Det kommende afsnit indeholder en kort deskriptiv beskrivelse af dele af dataet.

5.3 Deskriptiv beskrivelse af dataet

I nedenstående afsnit fremgår forskellige nøglekarakteristika for datagrundlaget. Formålet er at give et overblik over fordelingen af variablene i datagrundlaget, inden dataet benyttes i modellerne.

5.3.1 Antallet af boligtyper og placeringen

I datagrundlaget er der enfamiliehuse, ejerlejligheder og rækkehuse. Antallet af hver fremgår af nedenstående tabel 4:

Tabel 4 - Boligtyper og antal

Boligtype	Antal	Procent af Total
Enfamiliehus	437 101	66.7%
Ejerlejlighed	120 616	18.4%
Rækkehus	97 846	14.9%
Samlet antal	655 563	

Af de tre typer er enfamiliehuse den mest udbredte i dataet, efterfulgt af ejerlejligheder og sidst rækkehuse. Dette hænger sammen med, at parcel- og villahuse er den mest almindelige ejendomstype i Danmark. Eftersom der senere i projektet konstrueres tre individuelle modeller, vil forskellen i antal

ikke være af betydning. I nedenstående figur 1 fremgår den geografiske placering af de respektive ejendomstyper.

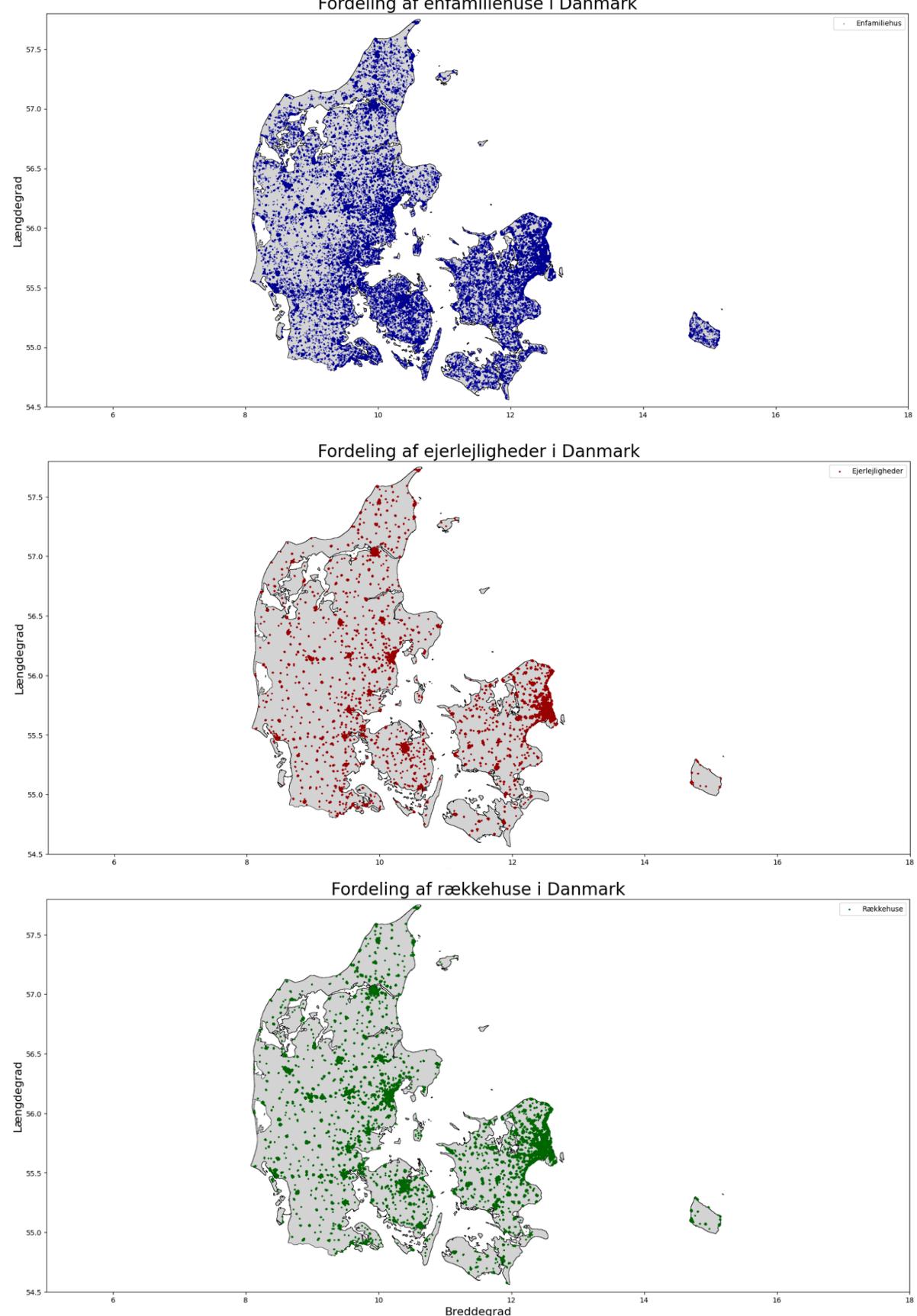
Af figur 1 kan det bemærkes, at enfamiliehuse dækker en stor del af Danmark, mens ejerlejligheder og rækkehuse synes at være placeret i grupperinger. Derudover fremstår enfamiliehus også som den boligtype, der er bedst repræsenteret i yderområderne, og der er derfor grund til at tro, at gennemsnitspriserne for referenceejendomme vil være mest nøjagtig for enfamiliehuse. Dette fremgår tillige af opdelingen efter region i tabel 5, hvor Nordjylland synes at være underrepræsenteret.

Tabel 5 - Fordeling af boligtype efter region

Boligtype	Region	% af samlede	Antal
Ejerlejlighed	Region Hovedstaden	36.23%	43705
	Region Midtjylland	23.38%	28198
	Region Syddanmark	21.40%	25811
	Region Nordjylland	10.63%	12823
	Region Sjælland	8.36%	10079
Enfamiliehus	Region Midtjylland	26.17%	114396
	Region Syddanmark	25.76%	112600
	Region Hovedstaden	18.51%	80899
	Region Sjælland	17.76%	77628
	Region Nordjylland	11.80%	51578
Rækkehus	Region Hovedstaden	37.69%	36880
	Region Midtjylland	21.01%	20560
	Region Syddanmark	17.61%	17233
	Region Sjælland	17.03%	16663
	Region Nordjylland	6.65%	6510

At nogle regioner er underrepræsenterede kan have indflydelse på modellens træfsikkerhed i disse regioner. Dels på baggrund af den mindre nøjagtige gennemsnitspris for området, men også hvis modellen generelt har svært ved at prissætte ejendomme i regioner uden det største grundlag. Dette afhænger dog af, hvor vigtig både gennemsnitsprisen og regioner er for ejendomsvurderingerne.

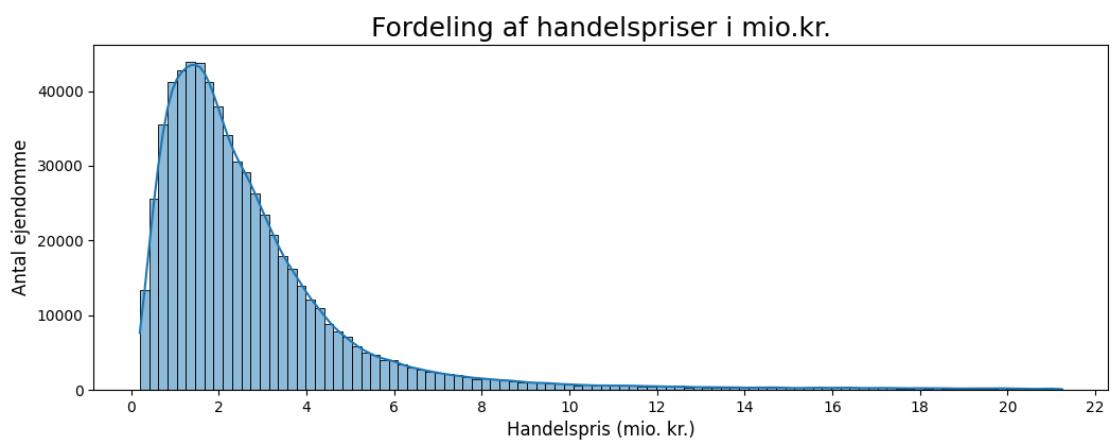
Figur 1 - Kort over placering af enfamiliehuse, ejerlejligheder og rækkehuse i datasæt



5.3.2 Fordeling af handelspriser

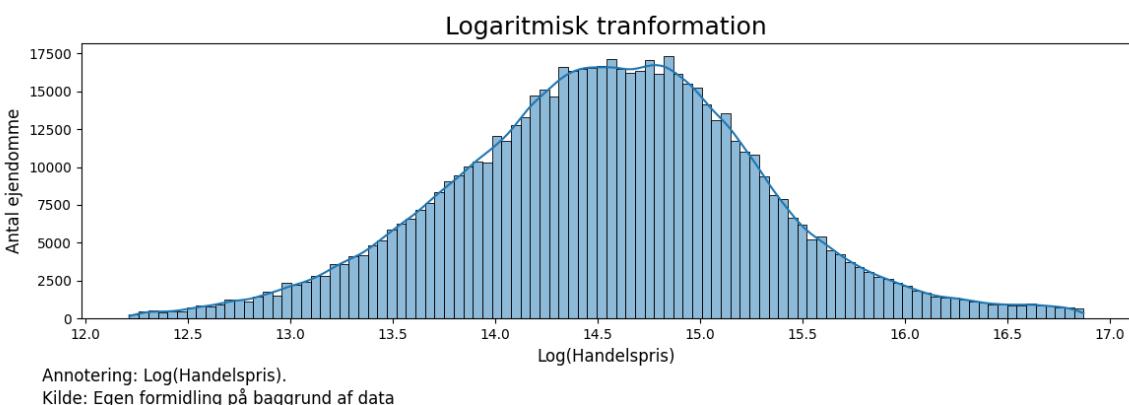
Fordeling af de realiserede handelspriserne, fremgår af nedenstående figur 2:

Figur 2 - Fordeling af handelspriser og logaritmisk transformation



Annotering: Bemærk at priserne er opgjort i mio.kr.

Kilde: Egen formidling på baggrund af data



Annotering: Log(Handelspris).

Kilde: Egen formidling på baggrund af data

Af figur 2 fremgår det, at handelspriserne er *right-skeewed*, og at en del ejendomme er solgt for over ti mio.kr., mens størstedelen er solgt for omkring 0,1 til fire mio.kr. Dette kan muligvis være et problem, såfremt projektets model er tilbøjelig til at bedømme lavere priser og dermed mindre træfsikker for højere priser. Løsningen på dette kan være at logtransformere handelspriserne ved brug af den naturlige logaritme, hvor fordelingen fremgår nederst i figur 2. Transformationen bevirket en mere symmetrisk fordeling, der kan afhjælpe de fornævnte problematikker. Dog medfører logtransformationen en problematik for gennemsigtigheden af de enkelte vurderinger, eftersom vurderingerne ikke vil kunne direkte sammenlignes i monetære værdier. Af denne årsag logtransformeres handelspriserne ikke. Ved afsnit 6.6 om modelopbygningen illustreres bivirkningerne ved ikke at logtransformere med afsæt i træfsikkerheden.

5.3.3 Distancemål

Fordelingen af distancerne mellem de enkelte ejendomme og forskellige interesseområder fremgår af nedenstående tabel 6.

Tabel 6 - Overblik over distancemål

Distance til:	Gennemsnit	Std	Min	25%	50%	75%	Maks
Kyst	9.394	11.014	0.000	1.660	5.294	12.812	65.522
Motorvej	11.767	24.878	0.018	2.204	4.582	12.067	261.032
Togspor	6.668	23.211	0.009	0.684	1.850	5.093	261.036
Lufthavn	51.415	36.272	0.958	22.173	47.341	74.723	283.744
Universitet	20.866	28.767	0.043	4.464	14.359	29.080	288.904
Skole	1.109	1.397	0.000	0.387	0.642	1.109	18.581
Børnehave	2.625	4.082	0.002	0.522	1.013	2.944	51.857
Vandløb/Sø	0.444	0.269	0.007	0.246	0.393	0.588	3.420
Skovområde	0.374	0.262	0.005	0.204	0.327	0.493	7.475

Der er relativ stor forskel på størrelserne af de forskellige mål. Eksempelvis er der for 25 pct. af fordelingerne mellem lufthavne og de øvrige variable en stor forskel på størrelsen. Dette fremgår desuden for de højeste distancer, hvor der er stor forskel på den højeste distance til vandløb og motorvej, togspor, lufthavn og universitet. Årsagen til dette er hovedsageligt antallet af de forskellige mål. Der er markant flere vandløb end lufthavne, og derfor er det tillige forventeligt, at der forekommer stor forskel i fordelingen. Det kan betyde, at modellen opfatter en større værdi som mere betydningsfuld, hvilket kan medføre, at lufthavn får tildelt en større betydning end eksempelvis vandløb. Derfor anvendes MinMaxScaler, der transformerer hver observation til en værdi mellem 0 og 1, hvor 0 angiver den laveste værdi inden for samme kategori og 1 angiver den højeste.

5.3.4 Øvrige variable

Af de øvrige variable er der flere kategoriske variable, der kræver omstrukturering inden de kan benyttes i modellerne. De kategoriske variable er; varmeinstallations-, tag- og vægmateriale typen, samt energimærkningen. Fordelingen af disse variable fremgår af nedenstående tabel:

Tabel 7 - Fordeling af udvalgte kategoriske variable

Varmeinstallation		Tag	
Type	Pct. af samlede	Type	Pct. af samlede
Fjernvarme/blokvarme	61.3	Fibercement asbest	32.9
Centralvarme	24.4	Tegl	28.8
Varmepumpe	9.9	Betontagsten	19.0
Elvarme	4.0	Tagpap hældning	6.8
Ovn	0.3	Tagpap	6.1
Blandet	0.1	Fibercement	2.2
Ingen varmeinstallation	≈0.0	Metal	2.1
Gasradiator	≈0.0	Andet materiale	1.1

Vægmateriale		Energimærkning	
Type	Pct. af samlede	Mærker (rang)	Pct. af samlede
Mursten	90.8	D (4)	29.1
Letbetonsten	2.9	C (5)	22.8
Træ	2.3	E (3)	16.8
Betonelementer	1.7	F (2)	9.4
Andet materiale	0.9	A (7)	9.3
Bindingsværk	0.9	B (6)	6.6
Fibercement herunder asbest	0.2	G (1)	6.0
Fibercement uden asbest	0.2		
Glas	0.1		
Metal	≈0.0		
Plastmaterialer	≈0.0		

Energimærket ranglister huses energiforbrug hvor A2020 er mest energivenlige energimærke, efterfulgt af A2015, A2010, B, C, D, E, F til G, som er mindst energivenligt. Disse er blevet omstruktureret til intervallet $A \rightarrow G$, da energimærkningen har ændret sig over tid (Energistyrelsen, 2023). Ligeledes transformeres de til numeriske værdier, der fremgår af parenteserne. Dette betyder at der fastholdes et rangerende hierarki mellem energimærkningen, således at en større værdi forbindes med højere energimærkning og forventes dermed at være positivt korreleret med handelsprisen.

For de øvrige kategoriske variable anvendes OneHotEncoder, hvilket medfører, at der skabes et antal kolonner for hver unik betegnelse i kategorierne. Det vil sige, at der eksempelvis skabes en kolonne for hver tagtype, der vil indeholde værdierne 0 og 1, alt efter om ejendommen har den pågældende tagtype eller ej. Det samme gør sig gældende for øvrige kategoriske variable, herunder sogn, region og byzone.

5.4 Metode

I de kommende underafsnit introduceres den valgte model med henblik på at beskrive modellens praktiske anvendelse. Endvidere introduceres de to statistiske evalueringsparametre som nærværende projekt benytter til at vurdere modellens præstation og træfsikkerhed. Sluttligt introduceres SHapley Additive exPlanations (SHAP), der har til formål at opfylde niveauet af gennemsigtighed, der jævnfør afsnit 5.2 er et kernelement af de nye ejendomsvurderinger.

5.4.1 Den valgte model

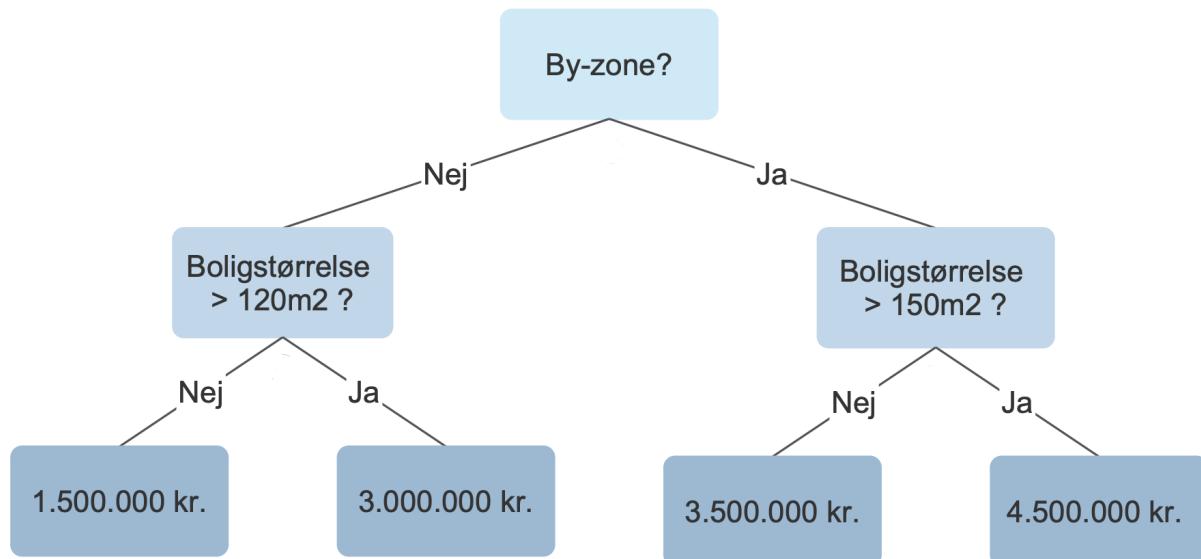
Der eksisterer to kategorier af superviserede maskinlærings (SML) modeller; regressions- og klassifikationsmodeller. Eftersom projektets model skal være i stand til at vurdere den numeriske pris af en ejendom, er det derfor nødvendigt at vælge en regressionsbaseret model. I nærværende projekt benyttes Extreme Gradient Boosting (XGBoost), der er en videreudviklet udgave af beslutningstræer. Årsagen til anvendelsen af XGBoost er, at det er en kraftfuld model, der i flere sammenhænge har vist at være overlegen sammenlignet med andre maskinlæringsalgoritmer (Shrestha, 2021). Det er endvidere også den empirisk fortrukne algoritme til besvarelse af lignende problemstillinger, som nærværende projekt berører (Hansen & Iversen, 2023, s. 59-81).

I de to næstkommende underafsnit introduceres beslutningstræer og XGBoost med henblik på at beskrive, hvordan XGBoost er en videreudviklet udgave af beslutningstræer.

5.4.2 Regressionstræer

Regressionstræer, der er en variant af beslutningstræer, er en type af SML-modeller, der er brugt til at forudsige en kontinuerlig variabel. Metoden har til formål at opdele dataet i flere regioner og tildele en konstant værdi til hver region – det er oftest middelværdien af variablen, der er fokus. Algoritmen opdeler datasættet i grupperinger, hvor der for hvert dybere niveau opdeles i to grupper. Opdelingen sker ved at minimere tabsfunktionen, som i regressionstræer oftest er de kvadrerede afvigelser. Denne rekursive opdeling af inputs fortsætter indtil algoritmen når et stop-kriterie, eksempelvis et minimum antal af observationer i hver opdeling eller en maksimal dybde. Disse kriterier er refereret til som hyperparametre, hvor der eksisterer en optimal værdi, der bedst forklarer dataet. Ved enden af træet, tages gennemsnittet af de boliger, der er i hver gruppe, hvilket giver den prædikerede værdi. Fordelen ved at bruge regressionstræer er, at de kan opfange ikke-lineære sammenhænge og er relativt intuitive (Vanderplas, 2016, s. 421-429). Dette er yderligere illustreret i nedenstående eksempel.

Tabel 8 - Eksempel på beslutningstræ⁴



Dette eksempel illustrerer, hvordan en kategorisk og numerisk inputvariabel benyttes til at opdele observationer. Første sektion, som indeholder alle observationer, opdeler algoritmen sættet baseret på, hvorvidt boligen er placeret i en by- eller landzone. I anden opdeling opdeler algoritmen observationerne i landzone baseret på, om de har et areal større end 120 kvadratmeter, mens boliger i byzone opdeles efter, om deres areal er større end 150 kvadratmeter. Sluttligt tages gennemsnittet af de boliger, der fremgår af de fire sidste opdelinger.

5.4.3 XGBoost

Modellen er udviklet af Tianqi Chen og Carlos Guestrin (2016), der formulerede en robust og effektiv udgave af beslutningstræer. XGBoost benytter sig af boosting, som er en teknik, der opstiller en udførlig model baseret på et antal af mindre kraftfulde modeller. Det vil sige, at en række af svage beslutningstræer kombineres i én samlet model, der oftest har høj træfsikkerhed. Modellen opstiller sekventielt de efterfølgende træer på baggrund af det foregående træs afvigelser. Det vil sige, at første træ udføres som beskrevet i ovenstående afsnit, det andet træ anvender afvigelserne af det første træ til at opstille nyt træ. Dette fortsætter sekventielt, hvor modellen overordnede set forbedres ved at tilpasse sig afvigelserne fra det foregående træ. I praksis vil denne rekursive proces fortsætte indtil modellen har forklaret hele datasættet eller andet kriterie opnås (Hansen & Iversen, 2023, s. 65-68). Denne rekursive boosting metode er refereret til som *Gradient Boosting*, der beskriver processen ved

⁴ Figuren er illustrativ og værdierne er hypotetiske.

at tilføje nye beslutningstræer til at forklare afvigelserne af de foregående træer. Gradient boosting involverer tre elementer;

1. Tabsfunktion

Tabsfunktionen beskriver afvigelsen mellem det estimerede og faktiske resultat. Typen af tabsfunktion afhænger af, hvorvidt det er regressions- eller klassifikationsvarianten, der anvendes. For regression bruges eksempelvis den gennemsnitlige kvadrerede afvigelse.

2. En weak learner

En weak learner er de tilføjede træer, der bruges til at forklare afvigelserne af det foregående træ og korrigerer dem.

3. Aggregering

De tilføjede træer tilføjes sekventielt ved brug af en gradient descent proces, der minimerer afvigelsen, når nye beslutningstræer tilføjes. Efter beregning af afvigelsen skal beslutningstræerne tilføjes, såfremt de reducerer den overordnede afvigelse.

5.4.4 Hyper-parameter tuning

Følgende hyperparamentre er udvalgt til at kontrollere XGBoost modellen:

- ***n_estimators***: Beskriver det samlede antal træer i modellen. Flere træer vil øge træfsikkerheden, men samtidig øge sandsynlighed for overfitted model og træningstiden.
 - Afprøvede værdier er [500 ; 800 ; 1000]
- ***learning_rate***: Beskriver størrelsen af skridt mod optimalt niveau. Lavere værdi betyder, at modellen langsomt bevæger sig mod optimum og reducerer sandsynligheden for, at forbipassere optimal løsning.
 - Afprøvede værdier er [0,005 ; 0,01 ; 0,05]
- ***max_depth***: Indebærer maksimal dybde af hvert beslutningstræ. Større dybde vil opfange mere komplekse forhold, men øger også sandsynligheden for overfitted model.
 - Afprøvede værdier er [15 ; 20 ; 25 ; 30]

- ***subsample***: Andelen af træningsdata, som benyttes ved hvert træ. Denne hyperparameter introducerer tilfældighed og reducerer sandsynligheden for overfitted model.
 - Afprøvede værdier er [0,8 ; 0,9 ; 1]
- ***gamma***: Er den mindste krævede reduktion i tabsfunktionen for at opdele en gren i beslutningstræet og hjælper med at reducere sandsynligheden for overfitted model.
 - Afprøvede værdier er [0,1 ; 0,2 ; 0,3]

I afsnit 6.6 om modelopbygningen fremgår resultaterne af kombinationen af ovenstående værdier, som minimerer den kvadrerede afvigelse mellem prædiktionerne og faktiske handelspriser i træningsdataet.

5.5 Afvejning mellem træfsikkerhed og gennemsigtighed

Som påpeget i tabel 1 er årsagerne til det gamle ejendomsvurderingssystems utilstrækkelighed, dets manglende gennemsigtighed og træfsikkerhed. Årsagen var det relativt høje niveau af subjektivitet, der i flere tilfælde kan være svær at forklare og som skaber uensartede vurderinger. Ligeledes er de nye ejendomsvurderinger blevet beskyldt for at være uigennemskuelige, mens det ligeledes i flere tilfælde har vidst sig at være langt fra de faktiske handelspriser (Bech-Nielsen, 2023).

I relation til nærværende projekts metode er det derfor nærliggende at forholde sig til afvejningen mellem træfsikkerhed og gennemsigtighed. *Træfsikkerhed* refererer til modellens evne til nøjagtigt at vurdere en specifik ejendom så tæt som muligt på den faktiske handelspris i fri handel. *Gennemsigtighed* referer i denne sammenhæng til, hvorledes den enkelte vurdering kan forklares med afsæt i objektive mål for den enkelte ejendom. Denne afvejning læner sig op ad den generelle forskel mellem Black- og white-box modeller, hvor black-box modeller er karakteriseret ved at være enormt træfsikre, men komplicerede og svære at forklare, mens white-box modeller er nemmere at forklare, men ikke lige så nøjagtige. XGBoost modellen kan i den forbindelse karakteriseres som en black-box model, da det er svært at gennemskue den overordnede model, når der anvendes et stort antal træer, som er sekventielt afhængige af hinanden.

Med afsæt i kritikken af det gamle vurderingsystem forekommer det ligeledes, at træfsikkerhed og gennemsigtighed er vigtige pejlemærker i udførelsen af det nye system. Afvejningen mellem disse er

dog svær at forene i samme model, men er en nødvendighed, når man som offentlig instans skal vurdere ejendomme til beskatningsgrundlag og ligeledes forklare vurderingen til den enkelte boligejer. Der eksisterer hertil redskaber til at øge gennemsigtigheden af Black-box modeller, herunder SHAP-værdier, som kan bruges til at øge gennemsigtigheden af individuelle vurderinger. I de næstkomende to under afsnit introduceres de evaluéringskriterier, som benyttes i projektet, efterfulgt af en introduktion til SHAP, som kan bruges til at opveje niveauet af gennemsigtighed.

5.5.1 Evaluéringsparametre

Til at vurdere nærværende projekts model tages der udgangspunkt i samme evaluéringsmål, som benyttes af Engbergudvalget, herunder den middel absolute difference/error (MAE) og Plus-minus 20 pct. (PM_{20}).

MAE beskriver den gennemsnitlige afvigelse i kroner ved formlen:

$$MAE = \frac{\sum_{n=0}^N |X_n - X_{realiseret_n}|}{N}$$

Hvor X_n er den prædikeret værdi for observation n , N er antallet af dataobservationer og $X_{realiseret_n}$ er den realiserede handelspris. MAE beskriver derfor den gennemsnitlige afvigelse mellem modellens forudsagte værdi og den faktiske værdi, opgjort i kroner. Dog kan store unikke afigelser for dyre ejendomme trække den gennemsnitlige afvigelse op (Skatteministeriet, 2014, s.107). Ligeledes vil der til træningen af XGBoost modellerne benyttes kvadrerede afigelser (MSE), der straffer større afigelser i højere grad og som på den måde kan være en bedre illustration af modellens afigelser.

PM₂₀ beskriver, hvor mange procent af ejendommene, som er prædikeret til at være indenfor ± 20 pct. af den realiserede handelspris. Udtrykket er beregnet ved følgende formel:

$$PM_{20} = \frac{1}{N} * \sum_{i=1}^N \mathbb{I}\left(\left|\frac{X_i - R_i}{R_i}\right| \leq 0,2\right) * 100$$

Hvor X_i er den prædikeret værdi, R_i er den realiserede værdi, N er det samlede antal observationer og $\mathbb{I}(\cdot)$ er en logisk indikator, der returnerer 1, hvis betingelsen er opfyldt, og 0 hvis ikke. Ved 20 pct. returnerer dette evaluéringsmål andelen af ejendomme, der er vurderet inden for Skatteministeriets

forsigtighedsprincip, som tager højde for den naturlige usikkerhed, der eksisterer i forbindelse med handelspriser for ejendomme (Skatteministeriet, 2016, s. 5).

5.5.2 SHAP-værdier og gennemsigtighed

SHapley Additive exPlanations (SHAP) er en metode til at få et indblik i Black-box modeller, der ellers er for komplicerede til at kunne forklares intuitivt. SHAP blev udviklet af Scott Lundberg og Su-in Lee (2017), med afsæt i Shaply værdier originalt formuleret af Lloyd Shaply (1953). *Shaply-værdier* er et koncept fra spilteori og benyttes i maskinlærings- og AI-sammenhænge til at beskrive den relative indflydelse fra hver variabel. I konteksten af nærværende projekts model kan SHAP derfor benyttes til at beskrive, hvilken indflydelse, som eksempelvis antallet af badeværelser og toiletter har for den specifikke ejendomsvurdering. Fremgangsmåden for SHAP er at betragte modellens output for en specifik ejendom, når en variabel er til stede, kontra når den ikke er. Forskellen er dermed den marginale indflydelse af den specifikke variabel. Dette udføres for forskelligartede kombinationer af variable, hvorved den gennemsnitlige indflydelse af hver variabel, på baggrund af alle kombinationer, kan fastsættes (Molnar, 2023). Overordnet kan SHAP benyttes til at opretholde det nødvendige niveau af gennemsigtighed for hver enkelt ejendomsvurdering ved at illustrere, hvor meget hver variabel betyder for den enkelte ejendomsvurdering.

5.6 Modelopbygning

Dette afsnit har til formål at gennemgå fremgangsmåden, der benyttes til at opstille projektets modeller. Dette involverer opdeling af datasæt i træning og test, iterativ eliminering af kolonner og hyper-parameter tuning ved krydsvalidering.

Step 1: Træning og testsæt

For at kunne evaluere træfsikkerheden af hver model er det nødvendigt at opdele datasættet i træning- og testsæt. Det vil sige, at modellerne trænes på én andel af datasættet og dernæst evalueres på baggrund af anden andel datasættet, som modellen ikke endnu har set. Til dette benyttes 80 pct. af datagrundlaget til træning, mens 20 pct. tilbageholdes til evaluering af modellerne. Observationerne, som indgår i hver andel, er valgt tilfældigt. For de tre datasæt er antallet af observationer derfor:

Tabel 9 - Træning- og testsæt størrelse

Rækkehuse		Ejerlejlighed		Enfamiliehus	
Træning	Test	Træning	Test	Træning	Test
78.276	19.570	96.493	24.123	349.680	87.420

Step 2: Iterativ eliminering af variable

Formålet med dette er skiftevist at eliminere variable og undersøge, hvordan dette påvirker modellens træfsikkerhed. Derfor fjernes hver variabel skiftevist, hvorefter modellerne trænes og testes på de respektive datasæt. Resultaterne heraf gennemgås ikke eksplisit, men denne iterative gennemgang viser, at elimineringen af sogn i gennemsnit øger træfsikkerheden med 0,5 - 1,5 pct. målt efter PM_{20} . Årsagen hertil er hovedsageligt, at der eksisterer 1916 forskellige sogne, hvilket øger dimensionen af dataet betydeligt som følge af OneHotEncodingen. Derfor fjernes variablen sogn for alle modeller.

Step 3: Tuning af hyper-parametre

Hver kombination af hyperparametre beskrevet i afsnit 6.4.4 gennemgås og rangeres dernæst efter træfsikkerhed. Hertil er resultaterne ens for alle modeller og fremgår af nedenstående tabel:

Tabel 10 - Resultat af hyperparameter tuning

	n_estimators	Learning_rate	Max_depth	Subsample	Gamma
Ens for alle modeller	1000	0.01	20	0.8	0.1

Som det fremgår af hyperparametrene, opnår alle modeller den mindste afvigelse ved at benytte de samme hyperparametre. Der arbejdes derfor ud fra disse hyper-parametre for hver model.

6 Resultater

Som beskrevet i afsnit 6 opstiller nærværende projekt én model for hver boligtype. Dette afsnit vil derfor indledningsvist evaluere de tre modellers træfsikkerhed og sammenligne med Engbergudvalgets modeller. Sluttligt vil det, baseret på én af projektets modeller, illustreres, hvordan gennemsigtighed kan repræsenteres ved brug af SHAP-værdier.

6.1 Evaluering og sammenligning med Engbergudvalgets modeller

Modellerne er hver blevet testet med evaluatingsparametrene beskrevet i afsnit 6.5.1, der indebærer MAE og PM_{20} . Afsnittet vil sammenligne projektets modeller med Engbergudvalgets modeller, samt så vidt muligt det fremgår af udvalgets rapport, også det gamle ejendomssystem. Ligeledes transformeres handelspriserne ikke logaritmisk, da dette ville medføre, at SHAP-værdierne senere i projektet tillige ville have en logaritmisk form, hvilket gør det mindre gennemskueligt. En logaritmisk transformation af handelspriserne vil dog jf. appendiks 11.3 have bidraget med en gennemsnitlig 0,7 pct. stigning i træfsikkerheden.

6.1.1 Evaluering af enfamiliehusmodellen

Evalueringen af modellerne i projektet, Engbergudvalget og det gamle ejendomsvurderingssystem fremgår af nedenstående tabel for enfamiliehuse:

Tabel 11 - Sammenligning af modeller for enfamiliehuse

Model	MAE (kr.)	PM ₂₀ (%)
Enfamiliehus model	502.719	58.0
Engbergudvalgets prototype for enfamiliehuse	298.000	67.5
Gamle ejendomssystem for enfamiliehuse	343.000	62.9

Annotering: I beregning af projektets træfsikkerhed, er det benyttet krydsvalidering.

Kilde: (Skatteministeriet, 2014, s. 108)

Enfamiliehuse er generelt den mindst træfsikre model på tværs af alle modeller. Dette hænger sandsynligvis sammen med, at det er den mest typiske boligtype i Danmark og derfor også har den største variation i handelspriserne. På tværs af ovenstående tre modeller fremstår modellen i nærværende projekt som den mindst træfsikre målt på PM_{20} og MAE. Projektets model for enfamiliehuse er henholdsvis 9,3 og 4,7 pct. mindre træfsikker end Engbergudvalget og det gamle ejendomssystem målt på PM_{20} . Endvidere er den gennemsnitlige afvigelse tilmed omrent 204 og 159 tusinde kr. højere end henholdsvis Engbergudvalget og gamle ejendomssystem. Dette vidner om, at nærværende projekts model, baseret på disse parametre, er mindre træfsikker og derfor ikke kan erstatte de to øvrige. Set i lyset af kompleksitetsforskellen mellem Engbergudvalget og projekts model, kan der ligeledes sættes spørgsmålstegn ved, hvorvidt den øgede kompleksitet og lavere gennemsigtighed er det værd, hvis ikke træfsikkerheden er bedre. Det er dog værd at påpege, at evaluatingsparametrene for Engbergudvalget og det gamle ejendomssystem er baseret på

træfsikkerheden for i alt 7.481 enfamiliehuse. I modsætning hertil er nærværende projekts model evalueret på i alt 87.420 enfamiliehuse. Dette kan være årsagen til den store afvigelse i MAE og ligeledes i PM_{20} mellem projektets model og de to øvrige, da grundlaget for evalueringen er markant større. På den anden side vidner dette også om, at projektets model er tilsvarende dårligere, da antallet af ejendomme vurderet udenfor ± 20 pct. er markant større.

Baseret på prisniveauet er der desuden nogle betydelige forskelle i træfsikkerheden mellem projektets og Engbergudvalgets model. Over- og underestimeringsprocenterne af begge fremgår af nedenstående tabel.

Tabel 12 - Over- og underestimering for enfamiliehuse

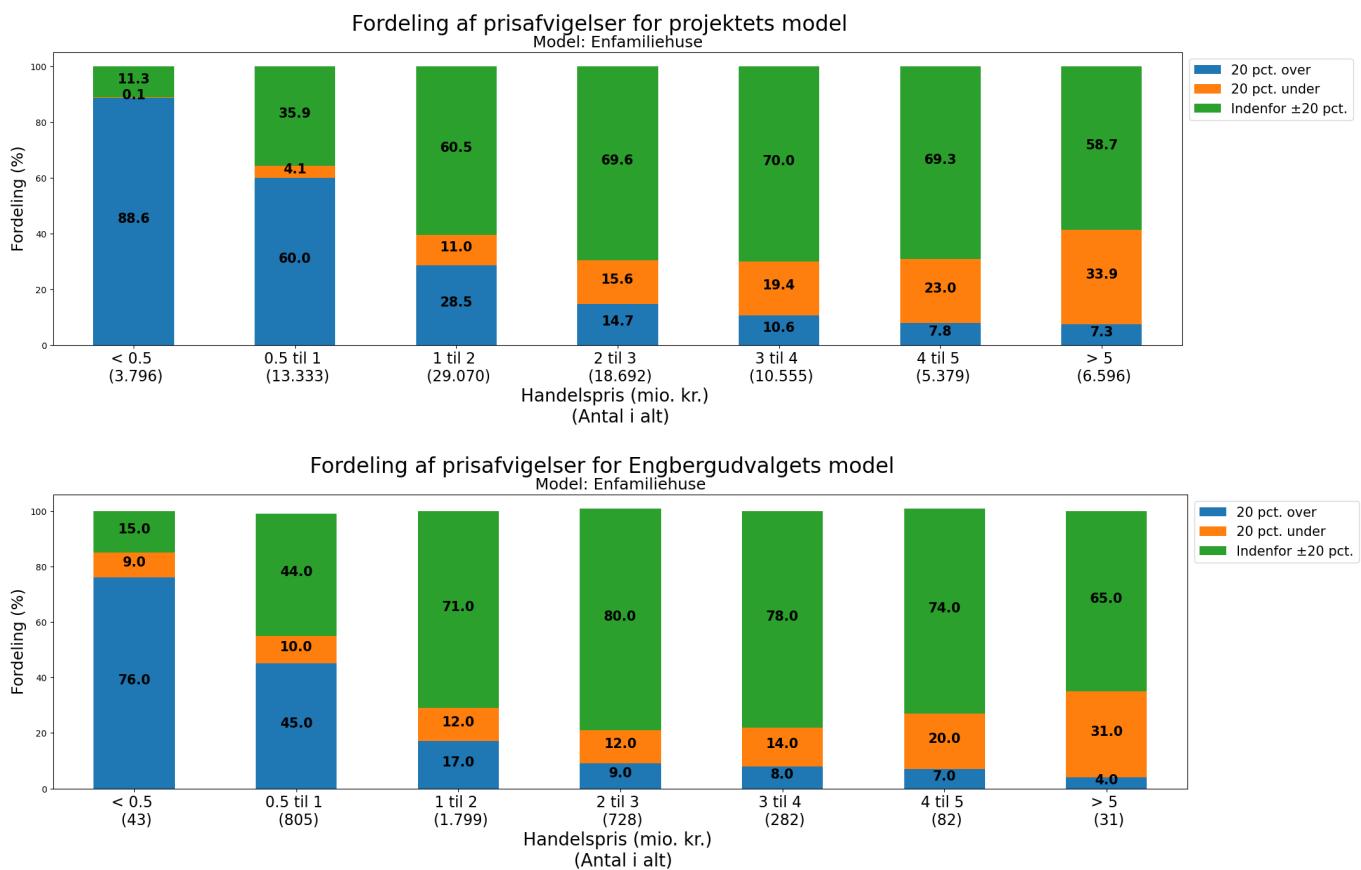
Model	PM ₂₀	Over 20 pct.	Under 20 pct.	Samlet
Projektets model for enfamiliehuse	58.0 pct.	27.9 pct.	14.0 pct.	100 pct.
Engbergudvalgets model for enfamiliehuse	67.5 pct.	20.2 pct.	12.3 pct.	100 pct.

Kilde: (Skatteministeriet, 2014, s. 114)

Tabellen illustrerer, at projektets model i større grad overestimerer ejendomsværdierne sammenlignet med Engbergudvalgets. Omtrent 7,7 pct. flere ejendomme overestimeres i projektets model, mens andelen af underestimerede ejendomme kun afviger med 1,7 pct. Dette afspejles ligeledes i nedenstående figur 3, der illustrerer andelen af PM_{20} , over- og underestimerede ejendomme efter den faktiske handelspris.

Figuren illustrerer, at begge modeller i et vist omfang overestimerer ejendomme med lavere handelspris og underestimerer dyrere ejendomme. Projektets model overestimerer dog billigere ejendomme i markant større omfang end Engbergudvalgets model, hvilket hovedsageligt forklarer den overordnede afvigelse i PM_{20} . Årsagen til dette kan skyldes, at datagrundlaget er større, hvilket sandsynligvis medfører en større variation, men samtidig også en større sandsynlighed for overfitting.

Figur 3 - Sammenligning mellem fordelingen af afvigelser for enfamiliehuse



6.1.2 Evaluering af rækkehushusmodellen

Projektets rækkehushus model sammenlignes med Engbergudvalgets og det gamle ejendomsvurderingssystem i nedenstående tabel 13.

Tabel 13 - Sammenligning af modeller for rækkehuse

Model	MAE (kr.)	PM ₂₀ (%)
Rækkehushus model	404.380 kr.	73.6 pct.
Engbergudvalgets prototype for Rækkehuse	223.000 kr.	81.0 pct.
Gamle ejendomssystem for enfamiliehuse	308.000 kr.	72.9 pct.

Annotering: I beregning af projektets træfsikkerhed, er det benyttet krydsvalidering.

Kilde: (Skatteministeriet, 2014, s. 108)

For rækkehushusmodellen tegner sig et lignende billede som ved modellen for enfamiliehuse. Projektets model for rækkehuse er betydeligt mindre træfsikker end udvalgets, men derimod marginalt bedre end det gamle ejendomssystem målet på PM₂₀. Modellen er 7,4 pct. mindre træfsikker end

Engbergudvalgets model og akkurat bedre end det gamle ejendomssystem ved en 0,7 pct. højere PM_{20} . Ligeledes fejlestimerer projektets model rækkehuse med i gennemsnit 404.380 kr., hvilket er henholdsvis 96.380 kr. og 181.380 kr. mere end det gamle ejendomssystem og udvalgets model. Dette afspejler, at projektets model gennemsnitligt er mindre træfsikker, men det vidner også om to bemærkelsesværdige pointer.

Engbergudvalget og det gamle ejendomssystem er begge evalueret på præcist samme udsnit af rækkehuse, hvilket betyder at deres evalueringsparametre er 1:1 sammenlignelige. På trods af at projektets model har en større træfsikkerhed ved PM_{20} , men en højere MAE, kan dette betyde, at der kan være større variation i projektets datagrundlag end i dataet benyttet af Skatteministeriet (2014). Havde Skatteministeriets (2014) datagrundlag været af samme størrelse som nærværende projekts kunne begge modellers evalueringsparametre have udvist lavere træfsikkerhed. Dette kan dog ikke vides med sikkerhed, men det er en plausibel mulighed, som Engbergudvalget ligeledes tager forbehold for (Skatteministeriet, 2014, s.119-121). Når projektets model ikke er indenfor ± 20 pct. af handelsprisen kan det også indikere at den er markant mindre træfsikker end såvel Engbergudvalgets model og det gamle ejendomssystem.

Sammenligning af træfsikkerheden baseret på handelsprisniveauet mellem projektets model og Engbergudvalgets fremgår af nedenstående tabel 14.

Tabel 14 - Over og underestimering for rækkehuse

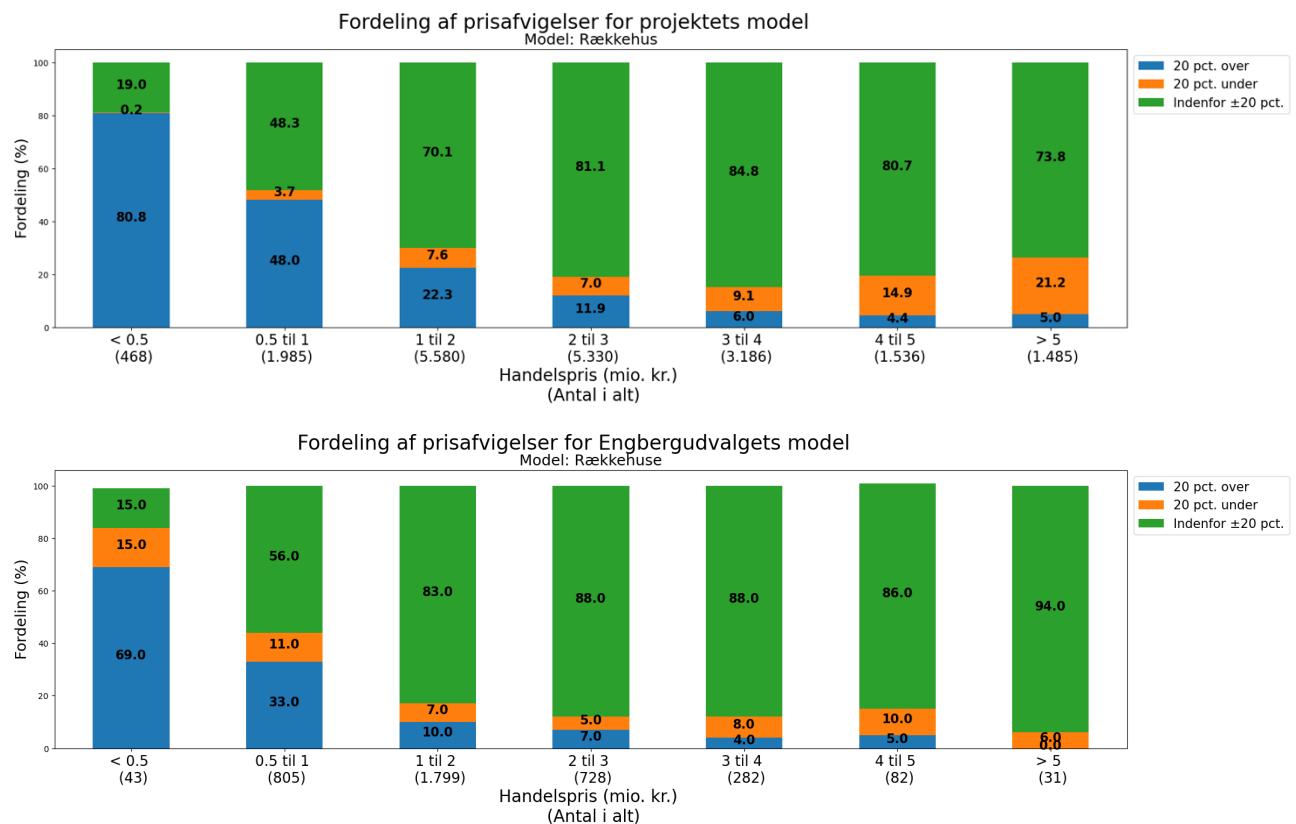
Model	PM_{20}	Over 20 pct.	Under 20 pct.	Samlet
Projektets model for rækkehuse	73.6 pct.	18.1 pct.	8.1 pct.	100 pct.
Engbergudvalgets model for rækkehuse	81.0 pct.	12.1 pct.	7.0 pct.	100 pct.

Kilde: (Skatteministeriet, 2014, s. 119)

Ligesom for projektets model for enfamiliehuse har modellen for rækkehuse ligeledes en tendens til at under- og overestimere i et større omfang end Engbergudvalgets model. Omfanget er dog mindre end tilfældet var for enfamiliehuse. For rækkehuse overestimerer projektets model 6 pct. flere rækkehuse med mere end 20 pct. over deres handelspris, mens der ligeledes underestimeres 1,1 pct. flere rækkehuse sammenlignet med Engbergudvalget. Ligeledes kan der sættes spørgsmålstegn ved validiteten af denne sammenligning i takt med, at Engbergudvalgets model evalueres på i alt 5.913 rækkehuse, mens nærværende projekt evaluerer træfsikkerheden for i alt 19.570 rækkehuse.

Over- og underestimeringen afspejles yderligere, når man sammenligner på tværs af de faktiske handelspriser. Dette fremgår af nedenstående figur 4.

Figur 4 - Sammenligning mellem fordelingen af afvigelser for rækkehuse



Af figuren fremgår et lignende scenarie, som udspillede sig ved modellen for enfamiliehuse. Modellen for rækkehuse har tilsvarende en tendens til at overestimere rækkehuse, som har en lav handelspris, mens der samtidig underestimeres ved høj handelspris. Tendensen til at underestimere dyre ejendomme er dog lavere end for enfamiliehuse blandt projektets modeller. Trods dette, har projektets model for rækkehuse en markant højere tendens til at underestimere dyre ejendomme, sammenlignet med Engbergudvalgets. Dette fremgår ved, at en øget handelspris i Engbergudvalgets model ser ud til at reducere andelen af underestimerede rækkehuse, mens det i projektets model ser ud til at øge andelen. Træfsikkerheden af Engbergudvalgets model for dyre ejendomme er kun baseret på 42 tilfælde mellem 4 og 5 mio.kr. og 35 tilfælde af ejendomme over 5 mio.kr. Hertil har nærværende projekt et betydeligt større datagrundlag på henholdsvis 1.536 og 1.485 observationer.

Hvorvidt der sammenlignes med Engbergudvalget eller alene på træfsikkerheden af projektets model, fremstår rækkehushusmodellen ikke tilfredsstillende. Projektets model er marginalt mere træfsikker end det gamle ejendomssystem, men stadigvæk mindre træfsikker end Engbergudvalgets model.

6.1.3 Evaluering af ejerlejlighedsmodellen

Evaluéringsparametrene for projektets model for ejerlejlighed, Engbergudvalgets og det gamle ejendomssystem fremgår af nedenstående tabel 15.

Tabel 15 - Sammenligning af modeller for ejerlejligheder

Model	MAE (kr.)	PM ₂₀ (%)
Ejerlejlighed model	422.414 kr.	78.9 pct.
Engbergudvalgets prototype for ejerlejligheder	192.000 kr.	83.4 pct.
Gamle ejendomssystem for enfamiliehuse	330.000 kr.	64.4 pct.

Annotering: I beregning af projektets træfsikkerhed, er det benyttet krydsvalidering.

Kilde: (Skatteministeriet, 2014, s. 108)

Træfsikkerheden for ejerlejligheder er den højeste af alle projektets modeller, hvilket tillige gør sig gældende for Engbergudvalgets model for ejerlejligheder. Projektets model for ejerlejligheder overgår dog stadigvæk ikke Engbergudvalgets model, som har en træfsikkerhed på 4,5 pct. mere. Dog er projektets model markant mere træfsikker end det gamle ejendomssystem med en træfsikkerhed på 14,5 pct. mere.

Ligesom tilfældet var for række- og enfamiliehuse forekommer afvigelsen i træfsikkerheden mellem projektets og Engbergudvalgets model i højere grad fra overestimering. Dette fremgår af nedenstående tabel 16.

Tabel 16 - Over og underestimering for ejerlejligheder

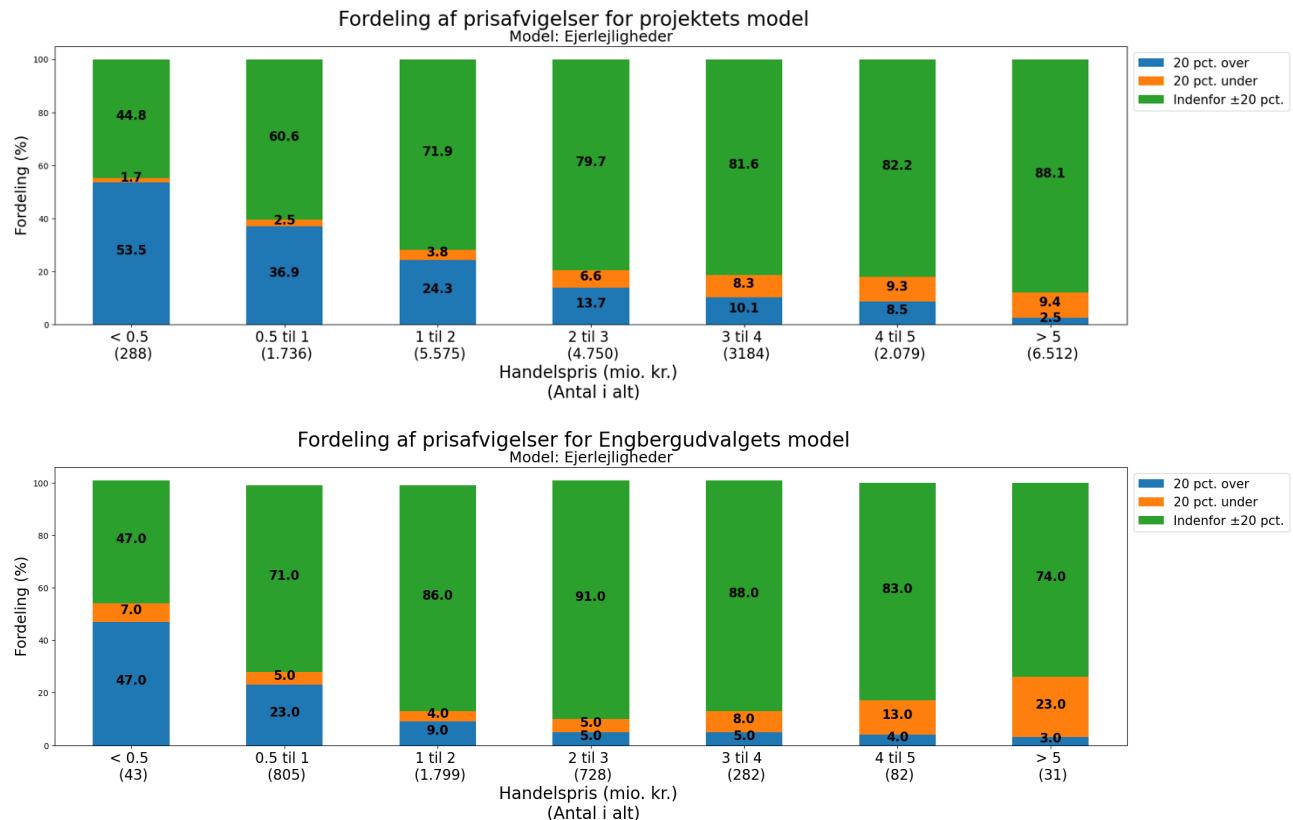
Model	PM ₂₀	Over 20 pct.	Under 20 pct.	Samlet
Projektets model for ejerlejligheder	78.9 pct.	14.3 pct.	6.8 pct.	100 pct.
Engbergudvalgets model for ejerlejligheder	83.4 pct.	11.2 pct.	5.3 pct.	100 pct.

Kilde: (Skatteministeriet, 2014, s. 122)

Af tabellen fremgår det, at projektets model for ejerlejligheder har en tendens til at overestimere ejendomspriserne. I alt overestimeres 3,1 pct. flere ejerlejligheder, mens der kun underestimeres 1,5 pct. flere ejendomme, når der sammenlignes med Engbergudvalget.

Baseret på de faktiske handelspriser minder under- og overestimerings mønsteret for ejerlejligheder om de to foregående modeller, eftersom modellen overestimerer lavere handelspriser og underestimerer ejendomme med højere handelspris. Dette fremgår af nedenstående figur 5.

Figur 5 - Sammenligning mellem fordelingen af afvigelser for ejerlejligheder



Projektets model for ejerlejligheder underestimerer ikke i samme omfang som de to foregående modeller for højere ejendomspriser. For ejendomme over 5 mio.kr. underestimeres blot 9,4 pct., hvilket i modellen for enfamiliehuse og rækkehuse var henholdsvis 33,9 og 21,2 pct. Dette hænger sandsynligvis sammen med at priserne for ejerlejligheder generelt er højere sammenlignet med enfamiliehuse og rækkehuse, hvilket tillige betyder, at modellen er trænet på datasæt med en større andel af ejendomme med højere priser.

For Engbergudvalgets model underestimeres tillige en større andel af ejerlejligheder over 5 mio.kr sammenlignet med projektets model. Dog har udvalgets model en større træfsikkerhed for de resterende priskategorier. For boliger over 5 mio.kr. har projektets model en 14,1 pct. større træfsikkerhed.

Overordnet fremstår Engbergudvalget model mere træfsikker end projektets model trods dennes mindre tendens til at overestimere ejerlejligheder.

6.1.4 Opsummerende for sammenligningen

Af sammenligningen mellem nærværende projekts modeller og Engbergudvalgets tydeliggøres det, at projektets model ikke er mere træfsikker. Hverken modellen for ejerlejligheder, række- eller enfamiliehuse har en højere PM_{20} end Engbergudvalgets respektive modeller.

Baseret på afsnit 6.5 om afvejningen mellem træfsikkerhed og gennemsigtighed kan det derfor virke besynderligt at gå på kompromis med gennemsigtigheden ved brug af XGBoost, når træfsikkerheden ikke forhøjes tilsvarende eller mere. Modellerne synes alle at have tendens til i højere grad at overestimere billige ejendomme, hvilket set over alle modeller, synes at forklare afvigelserne til Engbergudvalgets modeller. Det er gentagende gange i evalueringen blevet pointeret, at projektets tre modeller alle er blevet evalueret på et markant større datagrundlag, hvilket kan være årsagen til den til tider store forskel i træfsikkerhed i forhold til udvalgets modeller. Dette vil derfor indgå som element af projektets diskussion. Hvad end der sammenlignes med Engbergudvalgets modeller eller alene på den individuelle træfsikkerhed, fremstår projektets modeller ikke tilfredsstillende. Specielt modellen for enfamiliehuse, som har den laveste træfsikkerhed, er det problematisk, eftersom denne ejendomstype er den mest almindelige i Danmark. På baggrund af den manglende træfsikkerhed, vil der i første halvdel af næstkomende afsnit være fokus på at undersøge modellens prædiktioner mere i dybden, baseret på brugen af SHAP-værdier. Anden del af det kommende afsnit vil have til formål at undersøge, hvorvidt SHAP-værdierne kan tilfredsstille det nødvendige niveau af gennemsigtighed, som et offentligt vurderingssystem kræver.

6.2 SHAP-værdier og gennemsigtighed

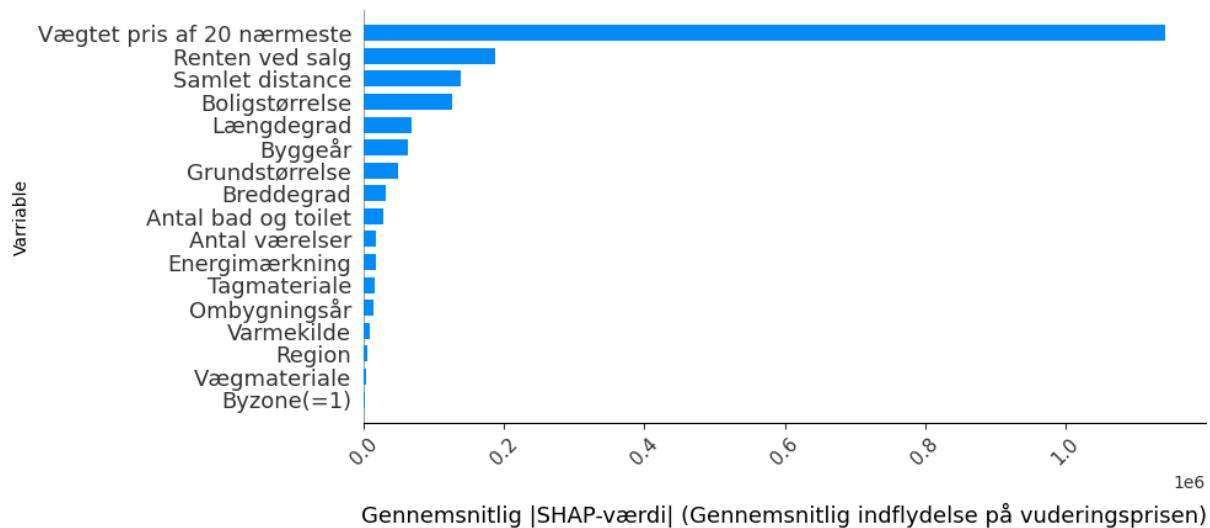
I afsnit 6.5.2 introduceredes SHAP-værdier med formålet at give indblik i, hvordan XGBoost modellen vurderer de specifikke ejendomme i projektets testsæt. I takt med evalueringen af projektets modeller blyses det, at træfsikkerheden generelt ikke er tilfredsstillende, hvorfor der i første omgang vil undersøges, hvordan de forskellige variable overordnet påvirker modellernes vurderinger. I undersøgelsen heraf tages der udgangspunkt i modellen for rækkehuse med henblik på at simplificere gennemgangen. Samme gennemgang for projektets øvrige modeller udføres løbende, men resultaterne heraf berøres blot i et begrænset omfang. Ydermere gennemgås der slutteligt to specifikke ejendommes vurderinger baseret på XGBoost modellen for rækkehuse. Disse har til formål at

illustrere, hvordan ejendomsvurderingerne kan se ud, når de er udført ved brug af XGBoost algoritmen, samt hvordan modellen vurderer delvist ens ejendomme forskelligt.

6.2.1 Global gennemsigtighed af modellen for rækkehuse

Ved at bruge SHAP-værdier er det muligt at undersøge, hvilke variable, der har den største betydning for modellens vurderinger. Nedenstående figur⁵ 6 viser sammenhængen mellem modellens variable og deres gennemsnitlige SHAP-værdi, som fortolkes ved variablenes gennemsnitlige indflydelse på modellens prædiktioner.

Figur 6 - Variable rangeret efter indflydelse på vurdering

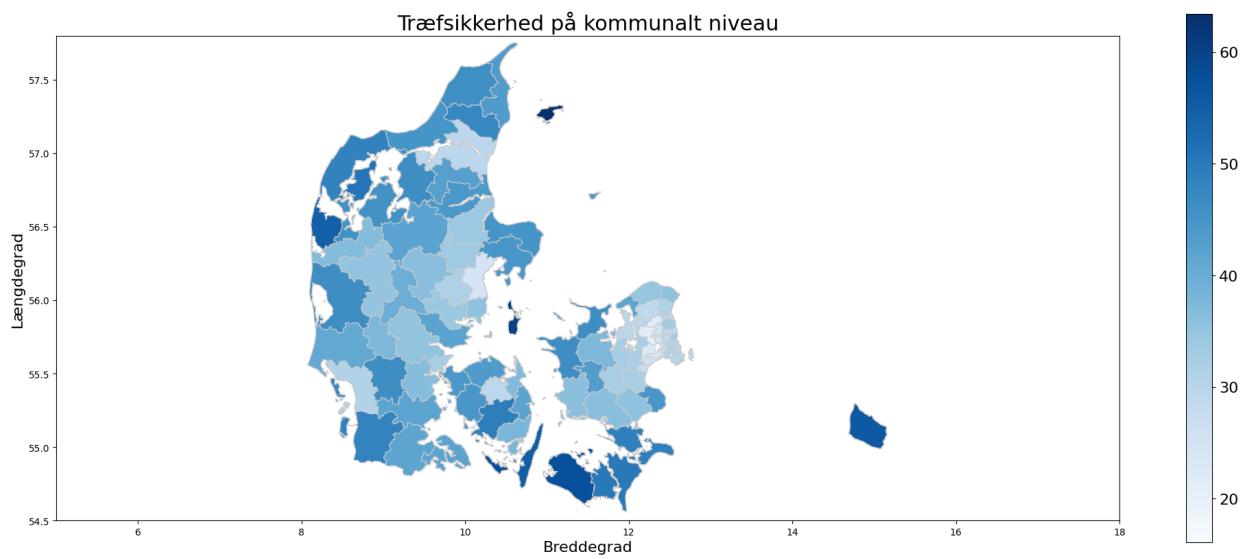


Af figuren fremgår det, at den vægtede pris af de 20 nærmeste ejendomme har den største betydning for modellens prædiktioner, efterfulgt af renten og distancemålene. Gennemsnitsprisen for de 20 ejendomme er knap seks gange mere betydningsfuld for modellens vurderinger end renten og distancemålene. Distancerne og renten fremstår dog betydningsfulde relativt til de øvrige, hvilket betyder, at projektets feature-engineering har været relevant at inkludere. Idet renten spiller en betydelig rolle kan det synes besynderligt, at renten ikke er inkluderet i det nye ejendomssystem eller Engbergudvalgets model (Skatteministeriet, 2014 og Skatteministeriet, 2016). Den tidligere beskrevet lave træfsikkerhed kan hænge sammen med, at modellen tilsyneladende finder de vægtede gennemsnitspriser yderst vigtig i bestemmelse af ejendomspriserne. Hvis denne har stor afvigelse fra

⁵ På baggrund af projektets anvendelse af OneHotEncoding, er flere af de kategoriske variable sammenlagt til én. Dette gælder ligeledes for distancemålene.

handelsprisen af den bestemte ejendom, vil modellen ligeledes fejlvurdere ejendommen ved at tillægge gennemsnitsprisen stor betydning. Dette kan desuden udledes af nedenstående figur, der viser træfsikkerheden på kommunalt niveau.

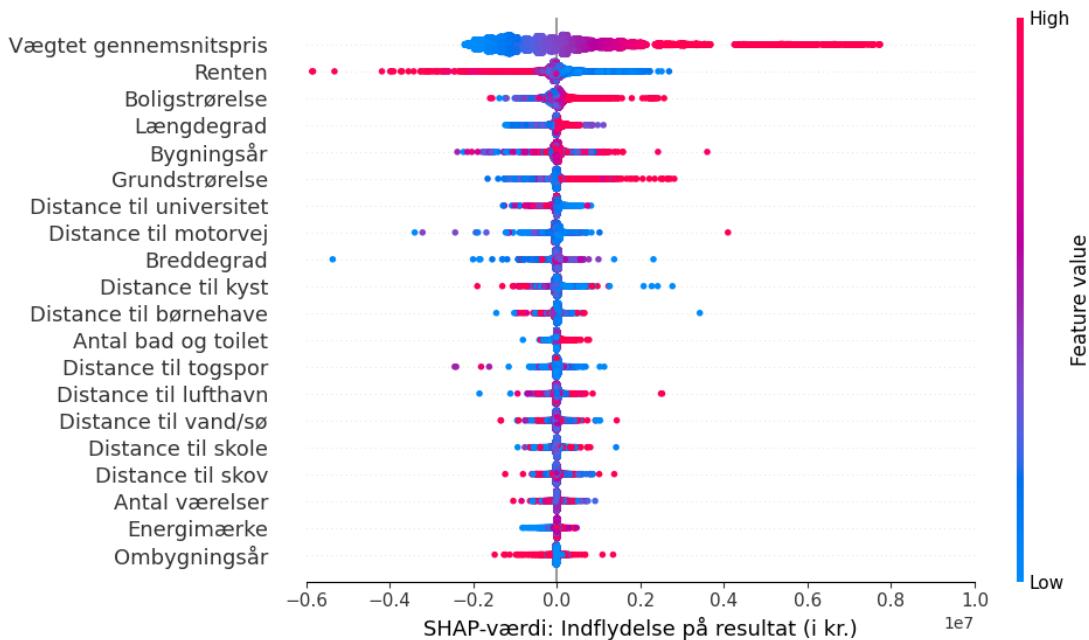
Figur 7 - Kort over træfsikkerhed for alle modeller på kommunalt niveau



Figuren viser, hvor andelen af alle ejendomstyper i testsættet, som afviger med mere end 20 pct. fra den faktiske handelspris, fordeler sig på kommunalt niveau. Sammenholdes denne figur med figur 1, fremgår det tydeligt, at træfsikkerheden er størst i de områder, hvor der er flest ejendomme og lavest i yderområderne. På denne måde afspejles en sammenhæng mellem antallet af ejendomme i nærområdet og træfsikkerheden, som opstår gennem de vægtede gennemsnitspriser betydning for modellerne. I et stort omfang synes dette at kunne forklare størstedelen af projektets lave træfsikkerhed. Dette vidner om, at der bør være stillede store krav til beregningen af den vægtet gennemsnitspris for de 20 nærmeste ejendomme, da den i stor grad påvirker træfsikkerheden og tillige har størst udfordringer i yderområderne.

For modellens øvrige numeriske variable fremgår der i nedenstående figur 8, hvordan disse påvirker modellens vurderinger.

Figur 8 - Numeriske variable og SHAP-værdi



Figuren illustrerer de enkelte SHAP-værdier for de numeriske variable i testsættet, hvor farverne fremhæver størrelsen af variablen. Af figuren aflæses, at flere af de inkluderede variable har den forventede indflydelse på ejendomsvurderingen af modellen. En højere rente kan eksempelvis spores til at have en negativ indflydelse på modellens prædiktioner, mens højere bolig- og grundstørrelse har en positiv effekt på den prædikeret værdi. Det samme gør sig ligeledes gældende med omvendt fortegn. For distancemålene er den egentlige sammenhæng tvetydigt. For flere af disse forekommer der ikke en direkte sammenhæng mellem større eller mindre distancer, og i hvilken retning det præger modellens prædiktioner. For enkelte af disse, såsom distancen til kyst og børnehave, synes der at være sammenhæng mellem, at lav distance medfører, at modellen vurderer ejendommen pris højere. Dette er foreneligt med de indledende forventninger og stemmer overens med projektets intuitive forståelse af sammenhængen.

6.2.2 Lokal gennemsigtighed af specifikke ejendomme

Formålet med dette afsnit er at undersøge, hvorvidt SHAP-værdierne kan opretholde niveauet af gennemsigtighed, hvis man anvender mere komplekse modeller til vurderingerne.

Til dette formål udvælges to delvist sammenlignelige boliger, hvor XGBoost modellens prædiktion af deres vurdering beskrives i detaljer. Den første ejendom vælges på baggrund af tilfældig indtastning, hvorefter der udføres Principal component analyse (PCA) til at reducere dimensionen af

testdataet til ti kolonner. Hertil findes den mest lignende ejendom udvalgt på Cosine-ligheden, der mäter vinklen mellem første og anden ejendoms vektorer.

I nedenstående tabel 17 fremgår de to ejendommens karakteristika, faktiske og prædikerede priser samt de enkelte variables indflydelse på modellens prædikerede værdi ved SHAP-værdier. Gennemsnitsprisen, som fremgår øverst i tabellen, er gennemsnitsprisen af alle de prædikerede værdier i testdataet. Herfra fratrækkes eller tillægges SHAP-værdierne og summeres slutteligt til den prædikerede værdi af modellen.

Tabel 17 - Beregningseksempel på to sammenlignelige boliger baseret på SHAP-værdier

Variabel	Bolig 1		Bolig 2	
	Værdi	SHAP	Værdi	SHAP
Gennemsnitspris		2.678.783 kr.		2.678.783 kr.
Antal værelser	5	- 33.193 kr.	4	- 5.558 kr.
Bygningsår	1969	- 25.528 kr.	1990	2.458 kr.
Boligstørrelse	100	- 44.681 kr.	91	- 130.956 kr.
Samlede distance til interesse mål	97.3 km.	34.708 kr.	143.7 km.	- 90.493 kr.
Længdegrad	10.81499	13.581 kr.	9.40303	- 105.698 kr.
Breddegrad	55.30855	- 52.023 kr.	56.43100	- 16.549 kr.
Grundstørrelse	1390	520.897 kr.	137	66.300 kr.
Varmekilde	Fjernvarme/blokvarme	18.495 kr.	Centralvarme	- 10.801 kr.
Renten ved salg	6.06	- 272.568 kr.	2.09	37.976 kr.
Tag-materiale	Fibercement asbest	- 31.762 kr.	Tegl	5048 kr.
Ombygningsår	2004	54.022 kr.	2008	91.662 kr.
Energimærke	B	60.134 kr.	D	- 12.041 kr.
Antal toilet og bad	2	- 107.990 kr.	2	- 5.573 kr.
Vægtede pris af 20 nærmeste	3.960.572	1.293.881 kr.	2.062.070	- 598.982 kr.
Øvrige variable		9.387 kr.		2.266 kr.
Bidrag fra variable:		1.437.360 kr.		- 770.941 kr.
XGBoost vurdering:		4.116.142 kr.		1.907.843 kr.
Faktisk handelspris:		4.191.690 kr.		1.527.030 kr.
Indenfor $\pm 20\%$?		Ja		Nej

Annotering: Samlede sum vil ikke nødvendigvis stemme overens, som følge af afrundinger

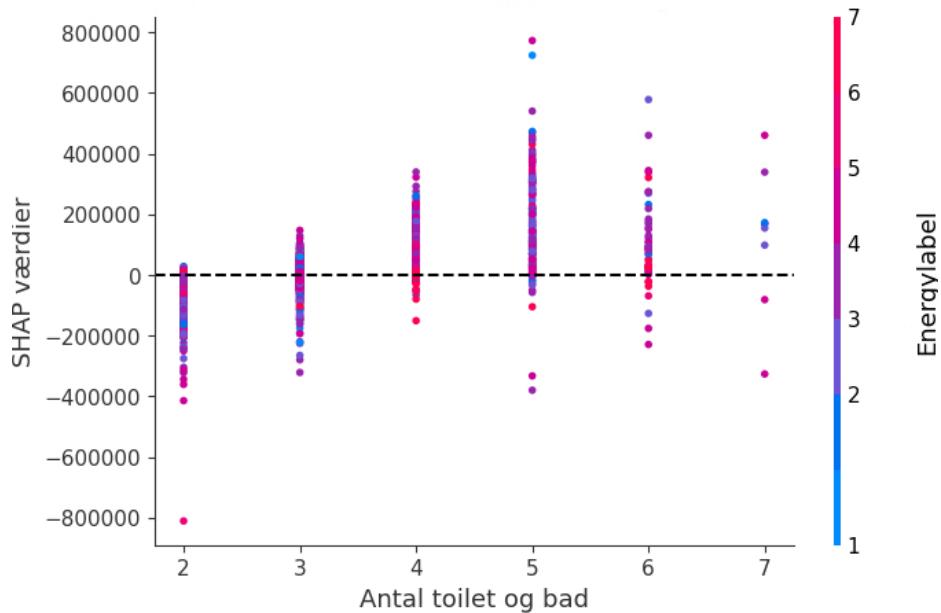
Kilde: Egne beregninger pba. SHAP-værdier

Tabellen er et illustrativt eksempel på, hvordan SHAP-værdier kan benyttes til at øge gennemsigtigheden af en kompleks model. De præcise værdier for de respektive variable har ingen direkte økonomisk betydning, forstået på den måde, at der eksempelvis ikke er beregnet værdien af ét værelse mere. I stedet skal værdierne fortolkes i sammenspil med den overordnede model, datasættet og de forudsætninger, som beskrives gennem projektet. Dette kan eksempelvis bemærkes ved antallet af bad og toilet, der er ens for de to ejendomme, men som i tilfældet af bolig 2 reducerer

værdien med kun 5.573 kr., mens værdien af bolig 1 reduceres med 107.990 kr. Dette skyldes dog det overordnede sammenspil mellem ejendommenes variable. Umiddelbart er det svært at gennemskue, hvorfor dette er tilfældet, da det kan skyldes sammenhænge mellem en lang række variable. I figur 9 synes der dog at være en sammenhæng mellem energimærkningen og SHAP-værdierne for antal badeværelser og toilet, hvor højere energimærke synes at forøge størrelsen af SHAP-værdierne. Dette kan være forklarende for, hvorfor bolig 1 oplever en større negativ værdi af antal toilet og bad sammenlignet med bolig 2.

Ydermere fremgår det også af tabellen, at den vægtede gennemsnitspris af de 20 nærmeste ejendomme har en betydelig stor indflydelse på den endelig vurdering. For bolig 1 tillægges en værdi på omrent 1.3 mio.kr. baseret på gennemsnitsprisen, mens bolig 2 fratages omrent 0.6 mio.kr. Dette afspejler for bolig 1 og 2 henholdsvis 41 og 45 pct. af den absolute værditilvirkning fra alle variable. Pointen fra den globale gennemgang underbygges derfor yderligere, da kroneværdien af de vægtede priser udgør en betydelig andel af alle tilskrivningerne.

Figur 9 - SHAP-værdier for antal badeværelser og toiletter efter energimærke



6.3 Opsummering af resultater

Som det fremgår af projektets resultater, er træfsikkerheden ikke tilfredsstillende for nogen af de opstillede modeller. Afvigelsen mellem projektets og Engbergudvalgets træfsikkerhed er størst for

enfamiliehuse, hvor projektets model i kun 58 pct. af testsættet formår at ramme indenfor ± 20 pct. af den faktiske handelspris. For rækkehuse og ejerlejligheden er afvigelserne mindre, men ikke bedre end Engbergudvalgets, hvor projektets træfsikkerhed er henholdsvis 73,6 og 78,9 pct. Afvigelserne beror i større omfang af overestimering af lavere handelspriser, som opstår i forbindelse med de vægtede gennemsnitsprisers betydning for modellens prædiktioner, hvor træfsikkerheden er lavest for de områder med færre referenceejendomme. Slutteligt er det illustreret, hvordan SHAP-værdier kan bruges til at forklare de enkelte prædiktioner for at opretholde det fornødne niveau af gennemsigtighed. Hvorvidt evalueringen mellem projektet og Engbergudvalget er retvisende, og om de enkelte vurderinger er tilstrækkelig gennemsigtigt, berøres fra flere vinkler i kommende afsnit.

7 Diskussion

Nærværende projekt har et gennemgående fokus på træfsikkerhed og gennemsigtighed i forbindelse med udviklingen af projektets modeller. Derfor vil projektets diskussion ligeledes berøre afvejningen mellem træfsikkerhed og gennemsigtighed med afsæt i resultaterne. Afsnittet omkring træfsikkerheden vil koncentrere sig om at adskille sammenligningsgrundlaget mellem projektet og Engbergudvalget, samt en kritik af de forskelligheder, der eksisterer.

7.1 Kan projektets og Engbergudvalgets modeller sammenlignes?

I beskrivelsen af projektets resultater fremlægges, at ingen af de udviklede modeller havde en tilstrækkelig høj træfsikkerhed sammenlignet med Engbergudvalget. Umiddelbart bør det af denne årsag fremstå tydeligt, at projektets modeller ikke er bedre end det formodede nye ejendomssystem. Alligevel bør der tages forbehold i denne sammenligning, da modellerne har vidt forskellige forudsætninger for at evaluere træfsikkerheden. I nedenstående tabel fremgår antallet af ejendomme i testsættet i projektets datagrundlag sammenlignet med Engbergudvalgets.

Tabel 18 - Sammenligning af testsæt

	Enfamiliehuse	Rækkehuse	Ejerlejligheder
Test antal i projekt	87.420	19.569	24.123
Test antal i Engbergudvalget	7.481	1.913	3.770

Projektets datagrundlag for evalueringen er omrent 12 gange større som Engbergudvalget for parcelhuse, ti gange større for rækkehuse og seks gange større for ejerlejligheder. Det er derfor ikke utænkeligt, at der blandt projektets testsæt er en større heterogenitet blandt ejendomme, hvilket kan reducere træfsikkerheden, såfremt modellen ikke opfanger dette. På den ene side er der tilsvarende større grundlag for at træne modellen, hvilket tillige bør gøre projektets model i stand til i højere grad at opfange denne heterogenitet. På den anden side er der ingen garanti for, at Engbergudvalgets modeller er skalerbare i den størrelsesorden, mens træfsikkerheden samtidig bevares. Det er ikke usandsynligt, at træfsikkerheden ved udvalgets model ville være lavere, hvis grundlaget for evalueringen havde været større. Med afsæt i dette kan der være forbehold for, at sammenligningen ikke fremlægger et retvisende billede af modellens præstation i forhold til det nye ejendomssystem. Sammenligner man i stedet med de foreløbige vurderinger, som DR benytter, gives et billede af træfsikkerheden i en mere sammenlignelig størrelsesorden. Ifølge DR er træfsikkerheden af det nye

system 64,1 pct. målt efter PM_{20} på baggrund af 82.603 ejendomme solgt i perioden 30. juni 2022 til 1. juli 2023 (Ingvorsen, Kielgast, Hecklen, & Ussing, 2023). Boligtypen af de 82 tusinde ejendomme oplyses dog ikke, men formodes at være en blanding af enfamiliehuse, rækkehuse og ejerlejligheder, da disse på tidspunktet af artiklen var de eneste, der havde fået vurderinger. Sammenlignet med den gennemsnitlige træfsikkerhed af projektets modeller, som er 70,2 pct. ved PM_{20} , fremstår projektets model derfor 6,1 pct. mere træfsikker end det nye ejendomssystem. Resultaterne, fremstillet af DR, repræsenterer derfor et bud på træfsikkerheden af det nye ejendomssystem, som er mere sammenligneligt med projektets evaluering. Denne sammenligning understøtter derfor pointen om, at det kan være misvisende at sammenligne træfsikkerheden mellem projektets modeller og Engbergudvalgets, da evalueringen ikke er baseret på et sammenligneligt grundlag.

I forlængelse af det mindre grundlag i Engbergudvalgets model, er der tillige være forskel i, hvordan nabopriserne beregnes. I Engbergudvalget beregnes nabopriserne ved handler i perioden 2006 til 2009, hvor mediankvadratmeterprisen for de 18 nærmeste for parcelhuse og ti nærmeste rækkehuse samt ejerlejligheder beregnes. I projektet er der jævnfør afsnit 6.2.1 anvendt en vægtning af gennemsnitsprisen baseret på kilometerafstanden mellem ens boligtyper i samme region. I forbindelse med, at projektets resultater har vist, at denne er den mest indflydelsesrige variabel i modellens prædiktioner, kan denne forbedres på flere måder. Foruden den alternative metode, der tidligere er beskrevet, kunne man tillige have vægtet afstanden og ligheden mellem de respektive ejendomme. Kvantificeringen af ligheden kunne eksempelvis være baseret på vinklen mellem de respektive ejendommes vektorer ved cosine-ligheden, som kort blev benyttet ved de lokale SHAP-værdier. En vægtning af gennemsnitsprisen på baggrund af distancen og ligheden, ville muligvis medvirke til en større træfsikkerhed i projektets modeller, da gennemsnitsprisen sandsynligvis vil forbedre repræsentationen af nærområdets priser. Der kunne i den forbindelse ligeledes være opstillet grænseværdier for, hvornår to ejendomme kan karakteriseres som sammenlignelige. Denne metode vil dog kræve en større mængde data end nærværende projektet har været i stand til at konstruere, samt bedre beregningsværktøjer. Det er uvist, hvorvidt dette havde øget træfsikkerheden af projektets modeller, men grundet betydningen af gennemsnitspriserne for området i modellerne vidner det om, at der bør stilles nøje krav til præcisionen i beregningen af denne. Det kan derfor ikke afgives, at projektets beregning af gennemsnitsprisen er årsag til den lave træfsikkerhed. Det er dog ligeledes en problematik, der er at finde i det nye vurderingssystem og kan derfor også være svaret på den kritik som systemet har mødt (Skatteministeriet, 2016, s. 13).

Ydermere påviser nærværende projekt, at der i det nye vurderingssystem er udeladt variable, der i projektet er bevist at have en stor betydning for modellens prædiktioner. Dette er illustreret ved 30-års renten for realkreditlån, der i projektets modeller er den næst vigtigste for modellens resultater. Hertil findes flere overordnede og generelle variable, der ikke inkluderes i det nye vurderingssystem, eksempelvis udbudsforhold, der fra et økonomisk perspektiv bør spille en rolle i værdien af en bolig (Danmarks Nationalbank, 2021, s. 3-7). Det kan derfor virke besynderligt, at rente- og udbudsforhold ikke er inkluderet i Engbergudvalgets model eller i udlægget fra Skatteministeriet om det nye ejendomsvurderingssystem (Skatteministeriet, 2014 & Skatteministeriet, 2016). Renten bør forventes at have en stor rolle i værdien af dyre ejendomme, da det øger den monetære omkostning forbundet med finansiering af boligkøb proportionalt med prisen af ejendommen. Ligeledes bør lokale udbudsforhold øge værdien af områdets ejendomme, da lavere udbud vil være forbundet med højere priser af alle ejendomme. Man kan således argumentere for, at der i vurderingen af en handelspris, som beror på en yderst kompleks sammenhæng mellem flere faktorer, bør være fokus på den specifikke ejendom. Derudover bør værdien af ejendomme ses i sammenspil med overordnede faktorer. Det nye vurderingssystem fremstår ikke sammenhængende, såfremt værdien af en ejendom ved bestemt tid, ikke sammenholdes med generelle økonomiske faktorer. Dette kan være en af årsagerne til den lave træfsikkerhed ved det nye ejendomssystems samt kritikken heraf, da projektet illustrerer, at renteforholdet er en væsentlig faktor at inkludere.

Overordnet illustreres væsentlige forbehold, der både understøtter projektets resultater og retter kritik mod Engbergudvalgets model og det nye ejendomssystem, i forbindelse med den overordnede problemstilling. Når dokumentationen og beregninger bag det nye ejendomssystem offentliggøres, bør det evalueres, hvorvidt deres model overkommer de problematikker, som kan identificeres i Engbergudvalgets model, såvel som i projektets udlæg. Endvidere bør der sættes store krav til beregningerne af den gennemsnitlige områdepris, hvor der muligvis kun bør sammenlignes delvist ens ejendomme i beregningen af denne, da en sådan problematik synes at reducere træfsikkerheden betydeligt i nærværende projekt.

7.2 Kan den fornødne gennemsigtighed opretholdes, hvis man benytter XGBoost?

Det gamle ejendomssystem blev suspenderet dels på baggrund af fejlagte vurderinger, men også den manglende gennemsigtighed i de enkelte vurderinger, som opstod i forbindelse med brug af subjektive vurderingsfaglige skøn. Tillige som beskrevet i afsnit 6.5, bliver maskinlærings algoritmer

oftest opfattet som sorte bokse, der beskriver modeltyper, som oftest er meget præcise, men hvor det ikke er direkte tydeligt hvordan modellen finder frem til den endelige vurdering. På baggrund af projektets resultater, er det umiddelbare resultat dog, at træfsikkerheden ikke har det ønskede niveau. Dog er det stadigvæk relevant at forholde sig til hvorvidt modeltypen som helhed kan benyttes til offentlige vurderinger, såfremt træfsikkerheden havde været tilfredsstillende.

Gennemsigtighed er i projektet refereret til, som sammenhængen mellem modellen og hvorvidt det er muligt præcist at forklare de endelige vurderinger, på en letforståelig måde. Ved brug af SHAP-værdier, bliver der i tabel 17 fremført to eksempler på hvordan modellen udfører vurderingerne. Baseret på denne, kan XGBoost modellen delvist opretholde det fornødne niveau af gennemsigtighed for den enkelte boligejer, der kan se hvordan den endelige vurdering er udmonstret. Med afsæt i en gennemsnitspris, tillægges eller fratrækkes ejendommen en vis værdi, alt efter ejendommens specifikke karakteristika. Intuitivt og letforståeligt, men måske alligevel ikke så gennemskueligt som Engbergudvalgets regneeksempel.

I Skatteministeriet (2014) fremføres et regneeksempel⁶ på hvordan specifikke vurderinger er lavet. Sammenlignes denne med nærværende projekts fremstilling i tabel 17, er der store ligheder, men også en væsentlig forskel. Af Skatteministeriets beregningseksempel, er der eksempelvis de samme parameterestimater for de samme variable. Trods at parameterestimaterne ikke bærer nogen økonomisk fortolkning, grundet korrelationen med nærområdeprisen, fremgår det stadigvæk tydeligere at en to ens boliger med 130 kvadratmeter, får samme værditillæg, da de ganges med samme faktor. I projektets beregningseksempel af tabel 17, kan der ikke opstilles samme forhold mellem variabelværdien og tillægget, da der eksisterer en flersidet sammenhæng med de øvrige variabelværdier. Dette blev illustreret ved at to ejendomme med samme antal toiletter og badeværelser, får forskellige tillæg i værdien. I projektet illustreres det, at denne forskel kan opstå på baggrund af energimærkningen, men at det samtidig er uvist i hvor stort omfang andre faktorer ligeledes spiller ind.

Denne manglende generalisering medfører, at det overordnede niveau af gennemsigtighed svækkes, fordi der ikke eksisterer direkte og generel sammenhæng mellem variabelværdien og dets tillæg til samlede vurdering. Som resultat af dette, vil en sammenligning mellem to ejendomsvurderinger,

⁶ Se appendiks 11.4

baseret på XGBoost, kunne give indtrykket af, at de to vurderinger er beregnede på forskelligvis, selvom det ikke er tilfældet. Samtidig kan det ligeledes være besværligt at forklare på en letforståelig måde overfor den enkelte boligejer hvorfor dette er tilfældet.

Brugen af SHAP-værdier bidrager dog med, at den enkelte ejendom præcist kan forklares i monetære værdier, hvordan bevægelsen mellem gennemsnitsprisen og den endelige ejendomsvurdering er udført, med afsæt i de enkelte variables indflydelse. Dette vurderes fra projektets side at være en tilpas høj gennemsigtighed, der medvirker til at afvejningen mellem træfsikkerhed og gennemsigtighed kan mindskes, ved brug af SHAP-værdier.

8 Konklusion

Nærværende projekt har en todelt målsætning om at undersøge, hvorvidt det er muligt at forbedre træfsikkerheden af det nye ejendomsvurderingssystem ved brug af maskinlæringsmodellen XGBoost, samt hvordan dette kan opnås uden at gå på kompromis med gennemsigtigheden af de enkelte vurderinger.

I tiden omkring projektets udførelse foreligger der ingen officiel dokumentation omkring det nye vurderingssystem, hvorfor det på baggrund af Skatteministeriets (2016) udlæg antages, at Engbergudvalgets (Skatteministeriet, 2014) rapport i overvejende grad efterligner det nye vurderingssystem. Med udgangspunkt i sammenligningen mellem projektets resultater og Engbergudvalget, er det umiddelbare svar på ovenstående spørgsmål, at projektets modeller ikke fremstår mere træfsikre en det nye vurderingssystem ved brug af maskinlæringsmodellen XGBoost. Projektets modeller for enfamiliehuse, rækkehuse og ejerlejligheder fremstår alle med lavere træfsikkerhed målt efter andelen af ejendomme prædikeret inden for ± 20 pct. Træfsikkerheden er lavest for enfamiliehuse og højest for ejerlejligheder, men alle fremførte modeller afviger i større eller mindre grad fra det nye vurderingssystem. Afvigelserne opstår i stort omfang grundet overestimering af ejendomme solgt for lavere beløb. Overestimeringen beror på vigtigheden af den vægtede gennemsnitspris for de 20 nærmeste ejendomme, der for projektets modeller, er den mest indflydelsesrige faktor for modellens prædiktioner. Beregningen af denne resulterer i, at modellens træfsikkerhed reduceres i takt med antallet af nærliggende ejendomme, hvilket betyder, at modellens træfsikkerhed er lavest i yderområderne, der er relativt mindre repræsenterede. I projektet argumenteres der for, at afvigelserne mellem det nye vurderingssystem og projektet kan skyldtes et uretmæssigt evalueringsgrundlag, da der for projektet er et markant større grundlag for evalueringen. Sammenlignes der i stedet med de foreløbige vurderinger i 2023 fra DR (Ingvorsen, E. S., Kielgast, N., Hecklen, A., & Ussing, J, 2023), der bidrager med mere præsentabelt evalueringsgrundlag, fremstår projektets generelle træfsikkerhed bedre end det nye vurderingssystem. Det nye vurderingssystem vurderes tillige at mangle en generel sammenhæng med overordnede faktorer, da nærværende projekt finder realkreditrenten specielt indflydelsesrig i modellens prædiktioner. Hertil kan kritikken af det nye vurderingssystem muligvis henføres til en generel manglende sammenhæng med øvrige faktorer, foruden en præcis tilgang til beregningen af nærområdets gennemsnitspris. Ydermere finder projektet belæg for, at såfremt et offentligt vurderingssystem udføres ved brug af komplekse maskinlæringsmodeller, kan det fornødne niveau af gennemsigtighed opretholdes, og de

enkelte ejendomsvurderinger kan forklares på en letforståelig og transparent måde. Af denne årsag kan der argumenteres for, at såfremt det offentlige forøger kompleksitetsniveauet af den valgte model, vil det stadig kunne lade sig gøre at forklare hver enkelt vurdering med afsæt i dets specifikke karakteristika. Det bør vurderes, hvorvidt det for den enkelte boligejer fremstår tydeligt, at forskellige karakteristika kan have varierende indflydelse på vurderingerne på tværs af ejendomme.

9 Bibliografi

Abildgren, K. (Maj 2017). A Chart & Data Book on the Monetary and Financial History of Denmark (Working Paper).

Adolfsen, J. F., Mønsted, B. M., Schmith, A. M., Martinello, A. T.-A., Gudiksen, S., & Sonberg, K. F. (2022). Segmentation of the Housing Market with Internet Data: Evidence from Denmark. s. 1-25.

Bech-Nielsen, P. C. (14. September 2023). Leder af Magtudredning: Skat skal lægge algoritmer bag nyt boligvurderingssystem frem for offentligheden, *RADAR*. Hentet 28. November 2023 fra <https://radar.dk/artikel/leder-af-magtudredning-skat-skal-laegge-algoritmer-bag-nyt-boligvurderingssystem-frem>

Chen, T., & Guestrin, C. (2016). *XGBoost: A Scalable Tree Boosting System*.

Danmark Nationalbank. (2021, Juni 21). *Robustheden på boligmarkedet bør styrkes*.

Danmarks Statistik. (December 2023). *Statistikbanken.dk*. Hentet fra <https://www.statistikbanken.dk/statbank5a/default.asp?w=1470>

DØRS. (2016). Dansk økonomi forår 2016. s. 215-309.

Enders, W. (2015). *Applied econometrics time series* (Årg. 4). Wiley.

Energistyrelsen. (2023). Det viser energimarkedet, hentet 10. December 2023 fra <https://ens.dk/ansvarsomraader/energimaerkning-af-bygninger/det-viser-energimaerket>

FinansDanmark. (2023). *Finansdanmark.dk*. Hentet fra <https://finansdanmark.dk/tal-og-data/boligstatistik/obligationsrenter/>

Hansen, J. Z., & Iversen, A. Ø. (2023). Prisen på ejerboliger 1992-2021. s. 59-82.

Hansen, J. Z., Iversen, A. Ø., & Stephensen, P. (2018). Ejeboliger i det 21. århundrede. s. 14-37.

Hecklen, A., Ingvorsen, E. S., & Kielgast, N. (17. November 2023). *DR.dk*, Skattevæsnet rammer forbi med 73 milliarder kroner: Se 95.000 skæve vurderinger, og hvor meget de er rettet. Hentet 17. 12 2023 fra <https://www.dr.dk/nyheder/penge/skattevaesnet-rammer-forbi-med-73-milliarder-kroner-se-95000-skaeve-vurderinger-og>

Hecklen, A., Ingvorsen, E. S., Ussing, J., Ørskov, O., & Nielsen, S. (31. Oktober 2023), Boligejere i tusindvis er fanget: Kan hverken få penge tilbage eller klage. Hentet 17. December 2023 fra DR.dk: <https://www.dr.dk/nyheder/penge/boligejere-i-titusindvis-er-fanget-kan-hverken-faa-penge-tilbage-eller-klage>

IGISMAP. (Marts 2022). Administrative Boundary Shapefiles – Regions, Municipalities, Postal Areas and More. IGISMAP.

- Ingvorsen, E. S., & Hecklen, A. (22. September 2023). Efter en række klager: Folketingets Ombudsmand går ind i sag om foreløbige ejendomsvurderinger. Hentet 14. December 2023 fra <https://www.dr.dk/nyheder/penge/efter-en-raekke-klager-folketingets-ombudsmand-gaar-ind-i-sag-om-foreloebige>
- Ingvorsen, E. S., & Hecklen, A. (11. Oktober 2023). *DR.dk*, Skattevæsnet lægger sig fladt ned efter undersøgelse af 3.000 pilskæve vurderinger. Hentet 16. December 2023 fra <https://www.dr.dk/nyheder/penge/skattevaesnet-laegger-sig-fladt-ned-efter-undersoegelse-af-3000-pilskaeve-vurderinger>
- Ingvorsen, E. S., & Hecklen, A. (2. November 2023). *DR.dk*, Skattevæsnet erkender nye, store forsinkelser af ejendomsvurderinger. Hentet 17. December 2023 fra <https://www.dr.dk/nyheder/penge/skattevaesnet-erkender-nye-store-forsinkelser-af-ejendomsvurderinger>
- Ingvorsen, E. S., Kielgast, N., Hecklen, A., & Ussing, J. (18. September 2023). *DR.dk*, 82.600 salgspriser afslører: Hver tredje ejendomsvurdering er meget skæv. Hentet 15. November 2023 fra <https://www.dr.dk/nyheder/penge/82600-salgspriser-afsloerer-hver-tredje-ejendomsvurdering-er-meget-skaev>
- Ingvorsen, E. S., Ussing, J., Hecklen, A., & Ørskov, O. (8. September 2023). *DR.dk*, Har kostet milliarder af kroner - nu viser interne papirer et system af 'forbløffende' lav kvalitet. Hentet 12. December 2023 fra <https://www.dr.dk/nyheder/penge/har-kostet-milliarder-af-kroner-nu-viser-interne-papirer-et-system-af-forbloeffende>
- Kettle, S. (5. Oktober 2017). *ESRI*, Distance on a sphere: The Haversine formula. Hentet 27. November 2023 fra <https://community.esri.com/t5/coordinate-reference-systems-blog/distance-on-a-sphere-the-haversine-formula/ba-p/902128>
- Kirkebæk-Johansson, L. (19. September 2023). *DR.dk*, Jakob købte sit hus for fem millioner. Fire måneder senere var det syv millioner værd ifølge skattevæsnet. Hentet 15. December 2023 fra <https://www.dr.dk/nyheder/penge/jakob-koehte-sit-hus-fem-millioner-fire-maaneder-efter-blev-det-vurderet-til-syv>
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. s. 1-10.
- Molnar, C. (21. August 2023). SHAP (SHapley Additive exPlanations). Hentet 15. December 2023 fra <https://christophm.github.io/interpretable-ml-book/shap.html>

- Ottensmann, J. R., Payton, S., & Man, J. (Januar 2018). Urban Location and Housing Prices within a Hedonic Model. *Regional analysis & policy*, 1(38), s. 19-35.
- Ravn, O. M., & Uhlig, H. (2002). On adjusting the Hodrick-Prescott filter for the frequency of observations. *The Review of Economics and Statistics*, 84(2), s. 371-380.
- Rigsrevisionen. (2013). Beretning til Statsrevisorerne om den offentlige ejendomsvurdering. s. 13-21.
- Rigsrevisionen. (2014). *Notat til Statsrevisorerne om beretning om den offentlige ejendomsvurdering*.
- Shapley, L. S. (18. Marts 1953). A Value for N-Person Games. s. 1-13.
- Shrestha, O. (14. September 2021). *C-SharpCorner*, XGBoost - The Choice Of Most Champions. Hentet 5. December 2023 fra <https://www.c-sharpcorner.com/article/xgboost-the-choice-of-champions/>
- Skatteministeriet. (30. August 2013). *Retsinformation.dk*, Vurderingsloven. Hentet 10. December 2023 fra <https://www.retsinformation.dk/eli/lta/2013/1067>
- Skatteministeriet. (2013). *Tillid til ejendomsvurderingerne*.
- Skatteministeriet. (2014). Forbedring af ejendomsvurderingen. s. 5-191.
- Skatteministeriet. (2016). Nye og mere retvisende ejendomsvurderinger. s. 1-13.
- Skatteministeriet. (2018). *Statistiske metoder til vurdering af grunde under ejerboliger*.
- Statsrevisorerne. (2021). Beretning om Skatteministeriets styring af det nye ejendomsvurderingssystem. s. 1-19.
- University of California, Berkeley. Museum of Vert, & Hijmans, R. J. (2015). Administrative and political divisions. *Administrative and political divisions*, 2.8. University of California, Berkeley. Museum of Vertebrate Zoology.
- Vanderplas, J. (2016). *Python Data Science Handbook* (1 udg.). O'Reilly Media, Inc.
- Vurderingsstyrelsen. (2023). *Vurderingsportalen.dk*, Ejendoms værdi for huse trin for trin. Hentet 10. December 2023 fra <https://www.vurderingsportalen.dk/ejerbolig/ejendomsvurdering/saadan-vurderer-vi-huse/saadan-fastsætter-vi-ejendomsvaerdien-for-huse/ejendomsvaerdi-for-huse-trin-for-trin>

10 Appendiks

10.1 Dataindsamlingsprocessen

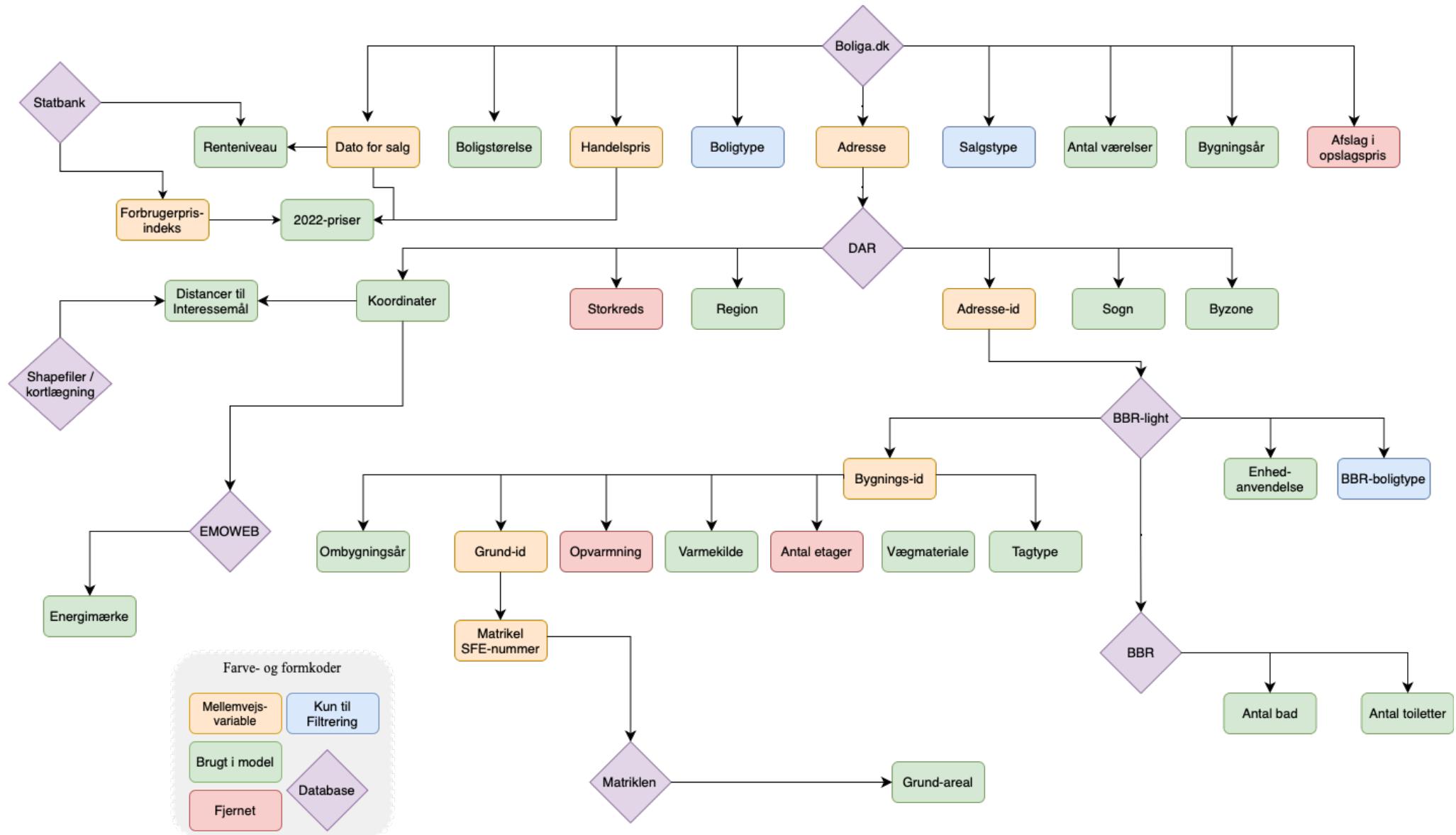
Som det er beskrevet i afsnit 6, har indsamlingen af det anvendte data, der er anvendt i projektet, været en udfordrende og kompliceret proces. Det har dog trods dette, været et vigtigt og nødvendigt element af projektets arbejde, da kvaliteten og størrelsen af dataet har været prioriteret.

Dataet stammer originalt fra Boliga.dk, der dagligt opgør nye solgte ejendomme. Herfra er der skrabet solgte ejendomme i perioden 1. januar 2000 til 20. november 2023, solgt ved alm. frit salg, ved at gennemgå HTML-koden af hver side i intervallet 1 til 22.883. Dataet hentet indebærer adresse, salgspris, antal værelser mm. Adresserne er derefter slæt op i Danmarks adresseregister (DAR), hvor der hentes adresse-id, koordinater, region mm. Adresse-id'et er benyttet til at søge i Bolig- og Bygningsregistre, herunder BBR-light og BBR, hvor der hentede forskellige karakteristika for enheden, boligen og grunden, herunder også matrikel-nummeret. Matrikelnummeret er herefter benyttede til at slå op i Matrikel databasen, hvor der opgøres den enkelte ejendoms grundareal.

Koordinaterne fra DAR er benyttet både i dataet, men også som opslagsværdi i EmoWeb (tidl. Diadem), hvor der er slæt den enkelte ejendom op indenfor en radius af 0,00018 km. (eller 0,18 m.) Koordinater bruges endvidere til at måle distancerne mellem ejendommen og forskellige interesseremål. Interesseremålene er hentet som Shapefiler, der i principippet er kortlægninger, som stammer fra Humanitarian OpenStreetMap (OSM), IGISMAP (IGIS) og University of California, Berkley. Distancerne er selv beregnet som beskrevet i dataafsnittet. Sluttligt er Danmarks statistik (STATBANK) brugt til at finde renteniveauet og forbrugerprisindekset, til at omskrive handelspriserne til 2022-priser, ved at benytte datoer for salget, som ligeledes bruges til at knytte renten til specifikke ejendomme.

Der er i denne proces brugt meget tid på detaljeret at skabe en pipeline til indsamlingen af dataet. Processen har været en iterativ proces, hvor der flere gange er vendt tilbage for at udvide datasættet yderligere. Der er benyttet adskillige API-løsninger og hentet mere data, som ikke nødvendigvis indgår i projektets modeller. Herunder eksempelvis antallet af etager, nedslag i pris, opvarmningsmiddel og storkreds. Antallet af etager og opvarmningsmiddel var ønsket at inddrage, men grundet et stort antal af manglende oplysninger for flere ejendomme, er disse ikke inkludere

Appendiks 10-1 - Dataindsamling grafik



10.2 Beskrivelse af variable i datagrundlaget

Appendiks 10-2 - Databeskrivelse

Variabel	Kilde	Detaljer
Handelspris	<i>Boliga</i>	Handelsprisen betalt for ejendommen, fremskrevet til 2022-prisniveau.
Antal værelser	<i>Boliga</i>	Antallet af værelser i boligen.
Opførselsår	<i>Boliga</i>	Året for boligens opførsel
Boligareal	<i>Boliga</i>	Areal til beboelse i kvadratmeter
Sogn	<i>DAR</i>	Den sogn boligen ligger i
Region	<i>DAR</i>	Den region boligen ligger i
Varmeinstallation	<i>BBR-light</i>	<p>BBR-oplysningskode til beskrivelse af typen af varmeinstallation:</p> <ul style="list-style-type: none"> 1 - Fjernvarme/Blokvarme 2 - Centralvarme 1 fyringsenhed 3 - Ovn til fast eller flydende brændsel. 5 - Varmepumpe 6 - Centralvarme 2 fyringsenheder. 7 - Elvarme 8 - Gasradiator 9 - Ingen varmeinstallation 99 - Blandet <p>BBR-koderne 2 og 6 vil være sammenlagt til én ved <i>Centralvarme</i>, da det vurderes ikke at være nødvendigt med sådan opdeling. Kode 6 var den type med færrest observationer.</p>
Byzone	<i>DAR</i>	Om boligen ligger i byzone (byzone=1) eller landzone (byzone=0).
Antal badeværelser	<i>BBR</i>	Antallet af badeværelser i boligen
Antal toiletter	<i>BBR</i>	Antallet af toiletter i boligen
Grundstørrelse	<i>Matriklen</i>	Antal kvadratmeter grund som boligen tilhører
Dato	<i>Boliga</i>	Dato for salget af bolig ved format (År-Måned-Dag)
Boligens anvendelse	<i>BBR-light</i>	<p>BBR-oplysningskode om boligens anvendelse. Listen er lang, men mest almindelige er:</p> <ul style="list-style-type: none"> 110 - Stuehus Landbrugsejendom 120 - Fritliggende enfamiliehus 130 - Række-, kæde- eller dobbelthus. 140 - Etageboligbebyggelse <p>I projektet anvendes terminologien enfamiliehus, villa eller parcel om kode 130, ejerlejlighed om kode 130 og</p>

rækkehus om kode 120. Kode 110 indgår ikke i datagrundlaget.

Energimærket på boligen. Energimærket ranglister huses energiforbrug, fra mest energivenligt til mindst ved: A2020, A2015, A2010, B, C, D, E, F og G. I projektet er først omskrevet til intervallet A til G, hvorefter disse er numerisk rangeret ved 7 til 1.

Energimærkning	<i>EMOWEB</i>	
Breddegrad	<i>DAR</i>	Boligens placering på kort. Koordinatsystemet anvendt er World Geodetic System 1984 (WGS84)
Længdegrad	<i>DAR</i>	
Distance til motorvej	<i>OSM</i>	Distancen er målt som kilometerafstanden mellem boligen og nærmeste motorvej, lufthavn, synlige togspor eller kyst i fugleflugtslinje. Der er anvendt koordinatsystemet WGS84 i overensstemmelse med længde- og breddegraderne.
Distance til togspor	<i>OSM</i>	
Distance til lufthavn	<i>OSM</i>	Af lufthavne indgår kun de store kommercielle lufthavne. For togspor indgår kun togspor over jorden og ikke metro. Kysten beskriver Danmarks administrative landegrænse uden grænsen til Tyskland.
Distance til skov	<i>OSM</i>	
Distance til vandløb, sø eller å	<i>OSM</i>	Skov og vandløb mm. er område-polygoner, hvor der er benyttet midtpunktet af disse til beregning af distancen.
Distance til skole	<i>OSM</i>	Distancen er udregnet ved kilometerafstanden mellem boligen og nærmeste skole, børnehave, eller universitet ved samme koordinatsystem som beskrevet ved længde- og breddegrader. Skoler involvere alle folkeskole, gymnasier, produktionsskoler og efterskoler mm.
Distance til Universitet	<i>OSM</i>	Universitet indebærer ligeledes også forskningshuse.
Distance til Børnehave og institution	<i>OSM</i>	
(Yder)Vægmateriale	<i>BBR-light</i>	<p>BBR-oplysningskode om boligens ydervægsmateriale.</p> <ul style="list-style-type: none"> • 1 - Mursten • 2 - Letbetonsten • 3 - Fiber cement herunder asbest <ul style="list-style-type: none"> • 4 - Bindingsværk • 5 - Træ • 6 - Betonelementer • 8 - Metal • 10 - Fiber cement uden asbest • 11 - Plastmaterialer <ul style="list-style-type: none"> • 12 - Glas • 80 - Ingen (Er fjernet i dataet) • 90 - Andet materiale
Tagmateriale	<i>BBR-light</i>	BBR-oplysningskode om boligens tagtype.

		<ul style="list-style-type: none"> • 1 - Tagpap med lille hældning • 2 - Tagpap med stor hældning • 3 - Fiber cement herunder asbest <ul style="list-style-type: none"> • 4 - Betontagsten • 5 - Tegl • 6 - Metal • 7 - Stråtag • 10 - Fiber cement uden asbest • 11 - Plastmaterialer <ul style="list-style-type: none"> • 12 - Glas • 20 - Levende tage • 80 - Ingen (Denne er fjernet i dataet) • 90 - Andet materiale
Ombygningsår	<i>BBR-light</i>	Året for registrerer ombygning.
Gennemspris for nærmeste 20 nærmeste	<i>Egen beregning pba. Boliga handelspris</i>	Gennemsprisen for de 20 nærmeste indebærer blot et enkelt kriterie, i form af, at den specifikke ejendom og referenceejendom, er i samme region. Gennemsnittet beregnes ved at vægte referenceejendommens værdi efter distansen mellem den specifikke ejendom og referenceejendommen. Der er kun udvalgt de 20 nærmeste ejendomme, målt i fugleflugtsafstande. For matematisk beregning kan der refereres til afsnit 6.2.1.
30 års realkredit rente	<i>Finans Danmark</i>	Lange obligationsrente på ugentlig basis
Forbrugerprisindeks	<i>STATBANK og Kim Abildgren (2017)</i>	Forbrugerprisindekset (2022=100). Kodenavn til Danmarks statistik er: PRIS111. Kodenavn til Abildgren (2017) er: S032M

10.3 Logaritmisk transformation af handelspriserne

Appendiks 10-3 - Sammenligning af evalueringsparametre ved logaritmisk transformation

	Enfamiliehuse		Rækkehuse		Ejerlejligheder	
	MAE	PM_20	MAE	PM_20	MAE	PM_20
Logaritmisk form	485.951 kr.	59.5	403.734 kr.	73.6	457.380 kr.	79.2
I monetære værdi	500.763 kr.	58.1	402.268 kr.	73.2	429,188	78.9

10.4 Regneeksempel på to ejendomme ved brug af Engbergudvalgets model

Appendiks 10-4 - Beregningseksempel på to delvist ens ejendomme ved Engbergudvalgets model

Boks 8.3.3. Regneeksempler

Eksempel 1: Parcelhus i Brande/Ikast på 130 kvm. Huset er fiktivt, men med tilstræbte realistiske værdier, dog (for oversuelighedens skyld) med færre viste inputvariable, end den egentlige model har (de ikke-viste variable er summeret under 'bidrag fra øvrige variable').

Indhold	Værdi	Parameterværdi	Bidrag til beskatningsværdi (kr.)
Konstant (samme værdi for samtlige ejendomme)			902.184
Overordnet estimat af ejendommen ud fra nærområdeprisen	1.324.440 (130 * områdeprisen)	0,65	860.886
Nærområdeprisen	10.188	-8	-81.504
Kommune	Ikast-Brande	-103.243	-103.243
Grundstørrelse	1.057	74	78.218
Bebygget areal	130	-1.435	-186.550
Afstand til jernbane	1.281m	34.329	34.329
Afstand til kyst	44.187m	-53.517	-53.517
Afstand til højspændingsledninger	1.564m	-5.246	-5.246
Bidrag fra øvrige variable			13.926
Total			1.459.483

Eksempel 2: Parcelhus i Hillerød på 130 kvm. Huset er fiktivt, men med tilstræbte realistiske værdier, dog (for oversuelighedens skyld) med færre viste inputvariable, end den egentlige model har (de ikke-viste variable er summeret under 'bidrag fra øvrige variable').

Indhold	Værdi	Parameterværdi	Bidrag til beskatningsværdi (kr.)
Konstant (samme værdi for samtlige ejendomme)			902.184
Overordnet estimat af ejendommen ud fra nærområdeprisen	2.690.610	0,65	1.748.897
Nærområdeprisen	20.697	-8	-165.576
Kommune	Hillerød	95.134	95.134
Grundstørrelse	1041	74	77.034
Bebygget areal	130	-1.435	-186.550
Afstand til jernbane	3.200m	23.985	23.985
Afstand til kyst	1.711m	-54.721	-54.721
Afstand til højspændingsledninger	325m	-27.911	-27.911
Bidrag fra øvrige variable			971.335
Total			3.383.811

Anm.: Bidrag til den endelige pris beregnes afhængig af, om variabeltypen er en talvariabel (som fx areal) eller en kategori-variabel (som fx tagtype). For talvariablene ganges tallet med parameterværdien, mens kategorivariablen har en værdi knyttet til hver kode. Bemærk, at flere variable, der oprindeligt har været talvariable, er blevet grupperet til kategorivariable for at opfange ikke-lineære effekter.

Kilde: (Skatteministeriet, 2014, s. 133) Boks 8.3.3