Chapter 1

Part I

I

1.1 Beskrivelse af data

Alle boligejere i Danmark betaler en skat, ejendomsvÃęrdiskat, som er baseret pÃě vÃęrdien af deres ejendom. Dette vil sige hele ejendommen inkl. grunden som boligen ligger pÃě. For at kunne gÃżre dette laver den danske stat offentlige ejendomsvurderinger som disse skatter bliver baseret pÃě. Det er derfor vigtigt at disse vurderinger er retvisende og ikke mindst forklarbare, sÃěledes at en borger kan forstÃě hvilke parametre der ligger til grund for ejendomsvurderingen. Til dette project har jeg valgt at arbejde med anonymiseret data fra mit arbejde i udviklings- og forenklingsstyrrelsen, hvor jeg til dagligt arbejder med netop dette. DatasÃęttet bestÃěr af ejendomssalg fra en 6 Ãěrig periode. Ud over selve huspriserne bestÃěr data ogsÃě af en lang rÃękke attributter som beskriver karakteristika ved selve boligen. Det kan f.eks. vÃęre tagmateriale, boligens opfÃżrelsesÃěr, information om stÃżrrelsen af huset og grunden eller bbr koder som dÃękker over boligens anvendelse. Der ud over bestÃěr data ogsÃě af en lang rÃękke attributter som fortÃęller noget om hvor boligens beliggenhed. Det kan f.eks. vÃęre boligens koordinater eller information om afstanden til kyst og skov eller afstand til motorvej og jernbane. Data kommer fra en rÃękke forskellige registre og offentlige styrrelser som eks. BBR og Styrrelsen for Dataforsyning og Effektivisering.

Til dette projekt vil jeg overordnet set prÃżve at se hvor godt man kan forudsige ejendomsvÃęrdier ud fra salgspriserne fra en 6-Ãěrig periode.

Jeg vil med Principal Component Analysis fÃě et overblik over de data der er til rÃědighed og fÃě et visuelt overblik over attributterne. Herefter vil jeg med en unsupervised learning forsÃżge at gruppere det data jeg har til at generer yderlige attributer som kan indgÃě i modellen. Jeg vil her specifikt prÃżve at se om det er muligt at gruppere salgene i forskellige boligtyper. Jeg vil i samme omgang ogsÃě forsÃżge at frasorterer outliers i data med anomaly

detection. Herefter vil jeg med regressions model forsÃÿge at kaste lys over projektets overordnede problem ved at forsÃÿge at forudsige huspriserne ud fra salgspriser. I tilfÃęlde af at modellen ikke ikke kan komme med en god prÃędiktion af en given ejendom vil det vÃęre muligt at denne ejendom bliver manuelt vÃęrdiansat af en sagsbehandler. Jeg vil derfor til slut med en classifikation forsÃÿge at estimerer om en ejendom skal ud til manuel sagsbehandling baseret pÃě dens estimerede ejendomsvÃęrdi.

1.2 Detaljeret beskrivelse af data

Det salgsdata som jeg har valgt at arbejde med dÃękker i udgangspunktet 'r nrow(train)' observationer med 'r ncol(train)' attributter. Inden jeg gÃěr i gang med at kigge pÃě data har jeg valgt at lave en oprydning i data. Det har jeg gjort fordi mange af attributterne bliver i mit daglige arbejde brugt i forbindelse med imÃÿdekomme diverse forretningskrav. Desuden dÃękker observationerne mange forskellige typer af ejendomssalg. Det er en blanding af parcelhussalg, rÃękkehussalg, sommerhussalg, salg af ejerlejligheder mm. og ud fra et forretningsmÃęssigt perspektiv giver det ikke mening at trÃęne en model pÃě alle salg og ejendomstypen vil pÃěvirke salgsprisen. F.eks. vil der pÃě sommerhuse vÃęre restriktioner pÃě hvor meget om Ãěret man mÃě bo i sommerhuset og der kan vÃęre i sommerhusomrÃěder vÃęre andre regler for hvad man mÃě bruge sin grund til end der er i et parcelhusomrÃěde. Jeg har derfor ligeledes valgt at reducerer antallet af observationer sÃěledes at de kun dÃękker almindelige parcelhus. Dette er gjort ved kun at beholde alle de ejendomssalg, hvor ejendommen i BBR er registreret med enheds- og bygningsanvendelsen 120.