Housing Prices – DAT158 rapport

Fredrik Enes, Eirik Sangiorgi Brakstad og Kristoffer Fjeldstad Madsen, 15.11.2023

BESKRIV PROBLEMET

SCOPE

Prosjektets mål er å gi personer som ønsker å selge eller kjøpe hus til å finne den riktige prisen for huset basert på ulike parametere. “Business Metrics” vil bli målt i verdien dette produktet har for kundene våre. Dersom de tar produktet i bruk, vil kundene ha mulighet til å spare tid på å sette takster på hus, og de vil ha en ide om hva de skal taksere huset for før takseringen gjennomføres. I tillegg vil det redusere stress for de som venter på taksering, ettersom de allerede vet en cirka sum huset deres er verdt. Stakeholders i dette prosjektet er en “Product Manager” og en “Software Developer”. Det kreves ikke noe særlig mer til dette forholdet.

Tenativ tidslinje
Milepæl 1:
Definere problemet og rydde i datasettet med one-hot-encoding osv.

Milepæl 2:
Se etter korrelasjoner i koden. Se om det er kolonner som kan droppes og finne hvilke data som kan droppes eller endres på for å forbedre treffsikkerheten til modellen.

Milepæl 3:
Finne den beste regression metoden og gjøre hyperparameter tuning på den.
Ressursene som kreves er datasett å beregne på. Dersom datasettene er for små så vil det kreves enda mer data og noen som skal samle inn dataen. Hvis datasettet er stort nok så kreves det en utvikler og en “product manager” for å fullføre produktet.

Milepæl 4:
Finne en måte å presentere prosjektet på.

METRIKKER

Vi ønsker å oppnå minimum 90% i business metric for å kalle det en suksess. Det tilsier at vi vil treffe på 9/10 huspriser.
For å måle hvorvidt systemet/løsningen fungerer så benytter vi en treffsikkerhet ved hjelp av R2 score og RMSE («Root-Mean-Squared-Error»). Dette vil henge sammen med «business objective» vi har satt, hvor vi ønsker å bli målt på verdien denne tjenesten vil ha for de som skal benytte seg av den. For at denne løsningen skal ha verdi for kundene er den nødt til ha høy nok «business metric» så de vet de kan stole på tjenesten.

DATA

Vi skal bruke data for hus og salgsprisen for disse husene på hvordan de har gått tidligere. Dataene hentes fra Kaggle, og dersom vi trenger mer data så vil det være mulig å hente inn flere data fra register over salg av hus og husets egenskaper. Per nå så er det ca. 3000 rader med data tilgjengelig, hvorav halvparten er treningsdata og andre halvparten er testdata. Vi tror ikke det er behov for mer data enn dataen som er tilgjengelig for å oppnå resultatet vi ønsker.
Vi har behov for rensing av data. Årsaken til dette er at dataene stort sett er «objects» og for at vi skal kunne bruke noen av regression algoritmene på den så må vi bruke «one-hot-encoding» eller droppe en del data. «One-hot-encoding» er en del av «feature engineering».

MODELLERING

Vi ønsker å utforske flere regression modeller, men vi starter med å utforske “gradient boosting regressor”. Vi planlegger å se på baseline ytelse ut fra hva andre tidligere har fått til på Kaggle. Der er det en del med mer erfaring enn oss som sannsynligvis kommer til å ha bedre modeller, men hva andre har fått er en del av baseline-ytelsen. Vi planlegger å lage noen enkle modeller i starten uten å gjøre for mye optimaliseringer. For å sjekke “feature importance” så kommer vi til å bruke korrelasjon for å sjekke hvor vi kan droppe data som ikke er så viktige. Vi vil forsøke å fjerne «outliers» også ved å fremstille de grafisk også hente de ut hvis verdien ikke ligger innenfor ønskede verdier.

DEPLOYMENT

Modellen skal settes i drift gjennom en web-app fra rammeverket “Gradio” som brukes ved hjelp av Python. Prediksjonen skal brukes i denne applikasjonen for å vise hvor mye man kan forvente at huset sitt er verdt basert på parametere.
Systemet må nok forbedres og vedlikeholdes etter det har blitt tilgjengeliggjort ettersom prisene på hus endrer seg regelmessig. Modellen må derfor trenes jevnlig på nyere data og da gjelder det å få tilgang til disse dataene.

KONKLUSJON

Gradio appen ble ikke like bra som vi forutså på grunn av en del verdier som måtte one-hot-«encodes» og det ble mange kolonner bortover. Vi fikk heller ikke det resultatet på score vi ønsket grunnet «encodingen» osv. så modellen ble ikke like effektiv som vi ønsket.