

# Prosjektrapport: Boligprisestimering og Prisprognoser

## Introduksjon

Formålet med dette prosjektet er å utvikle en tjeneste for boligprisestimering og predikering av fremtidige boligpriser. Tjenesten skal gi brukere et pålitelig estimat for dagens boligverdi og en prognose for fremtidig prisutvikling uten krav om innlogging eller deling av sensitive opplysninger. Prosjektet retter seg mot privatpersoner som søker informasjon om boligverdi og profesjonelle aktører som eiendomsmeglere og investorer som ønsker å ta mer informerte beslutninger.

## Business Objectives

Tjenesten er designet for å møte flere spesifikke forretningsmål:

- **Bedre investeringsbeslutninger:** Ved hjelp av prisestimering og prognoser kan eiendomsutviklere og investorer identifisere eiendommer og områder med høy avkastning, noe som gjør investeringsbeslutningene mer datadrevne og målrettede.
- **Forbedre prissettingsstrategi:** Tjenesten gir grunnlag for realistisk og konkurransedyktig prising av boliger, noe som gir eiendomsmeglere og boligselgere mulighet til å fastsette priser i tråd med markedsforholdene.
- **Identifisere lønnsomme utviklingsområder:** Prisprognoser gir boligutviklere innsikt i hvor etterspørselen sannsynligvis vil øke, noe som bidrar til strategisk plassering og lønnsomhet i nye boligprosjekter.

## Business Impact

Tjenesten har en betydelig forretningsmessig påvirkning:

- **Bedre beslutningsgrunnlag for investorer:** Estimering av nåværende og fremtidig verdi gir eiendomsutviklere og investorer et pålitelig verktøy for vurdering av avkastningspotensialet i ulike områder.
- **Forbedret prissetting for meglere og selgere:** Nøyaktige prisestimerer gir et solid grunnlag for realistisk prising, som igjen bidrar til raskere salg og konkurransedyktig markedsposisjon.
- **Støtte til boligutviklere:** Prediksjon av boligpriser gir verdifull innsikt i forventet etterspørsel, som hjelper til med planlegging og optimal ressursbruk i nye prosjekter.

## Sammenligning med tilsvarende Løsninger

Tilsvarende løsninger krever ofte innlogging med BankID og tilgang til personlige opplysninger, noe som kan virke unødvendig omfattende for brukere som kun ønsker et enkelt estimat og en prisprognose. Vår tjeneste gir brukerne den samme innsikten uten kompliserte krav, noe som gjør løsningen mer brukervennlig og tilgjengelig.

## Manuell Metode for Prisestimering

Uten maskinlæring ville en prisestimering måtte utføres manuelt på følgende måte:

- **Enkel input:** Brukeren fyller inn informasjon om boligen, som adresse, størrelse, og antall rom.
- **Resultatvisning:**
  - **Estimert boligpris:** Viser et estimat av boligens nåverdi.
  - **Prisprognose:** Viser en graf over forventet prisutvikling.
  - **Tilleggsinformasjon:** Viser gjennomsnittspriser i området og faktorer som påvirker prisene. Denne manuelle prosessen er tidkrevende og mindre nøyaktig enn en automatisert maskinlæringsmodell.

## Business Metrics for Ytelsesmåling

For å vurdere om systemet oppnår forretningsmålene, skal følgende business metrics følges:

- **Nøyaktighet på prisestimat:** Estimatenes bør være innenfor  $\pm 10\%$  av faktiske salgspriser i minst 80 % av tilfellene. Dette sikrer at modellen gir pålitelige resultater for brukerne.
- **Brukertilfredshet:** Brukere bør vurdere tjenesten med minst 4 av 5 stjerner i tilbakemeldinger, som indikerer at tjenesten oppleves som nyttig og brukervennlig.
- **Fullført brukersesjon:** Minst 70 % av brukerne som starter en prisvurdering, bør fullføre hele prosessen. Dette viser at nettsiden og estimeringsfunksjonaliteten er intuitiv og engasjerende.
- **Systemtilgjengelighet:** Nettsiden og estimeringstjenesten bør ha en opptid på minst 99 % for å sikre pålitelig tilgang for brukerne.

## Maskinlærings- og Software-Metrikker

Følgende tekniske metrikker vil brukes for å måle ytelsen til systemet og sikre at løsningene fungerer effektivt:

- **Treffsikkerhet (Accuracy):**
  - **Definisjon:** Andelen av prisestimater som faller innenfor et definert område rundt faktiske salgspriser, for eksempel  $\pm 10\%$  av virkelig verdi.
  - **Formål:** Måle modellens evne til å gi brukerne nøyaktige prisestimater.

- **Sammenheng med business objective:** Høy treffsikkerhet bidrar til brukertilfredshet og tillit til tjenesten, noe som er avgjørende for å oppnå minst 4 av 5 stjerner i brukervurderinger.
- **Kvadratisk avvik (Mean Squared Error, MSE):**
  - **Definisjon:** Gjennomsnittet av kvadratet av feilene mellom predikert og faktisk boligpris.
  - **Formål:** Gi et presist mål på hvor mye estimatene i gjennomsnitt avviker fra sanne verdier, som hjelper med å minimere feil.
  - **Sammenheng med business objective:** Lav MSE sikrer høy nøyaktighet, noe som øker brukertilfredshet og pålitelighet for tjenesten.
- **Latency (Forsinkelse):**
  - **Definisjon:** Tiden det tar for systemet å returnere et prisestimat etter at brukeren har sendt inn informasjon.
  - **Formål:** Måler systemets responsivitet, noe som påvirker brukeropplevelsen.
  - **Sammenheng med business objective:** Lav latency støtter høy fullføringsrate i brukersesjonene, ettersom brukere får raske svar og fullfører prosessen, noe som bidrar til målsettingen om minst 70 % fullføringsrate.
- **Throughput (Gjennomstrømming):**
  - **Definisjon:** Antall brukersesjoner eller prisestimer systemet kan håndtere per tidsenhet, for eksempel per minutt.
  - **Formål:** Måle systemets kapasitet til å håndtere flere brukere samtidig uten ytelsesproblemer.
  - **Sammenheng med business objective:** Høy throughput støtter systemtilgjengelighet og stabil ytelse, som bidrar til en oppetid på minst 99 %, noe som sikrer brukertilgang og tillit til tjenesten.

## Stakeholders

Prosjektet involverer flere viktige stakeholders:

- **Kunder/brukere:** Privatpersoner og profesjonelle som bruker tjenesten for å få estimer og prognoser for boliger. Forventningene er en enkel, nøyaktig tjeneste uten krav om sensitive opplysninger.
- **Eiendomsmeglere:** Profesjonelle brukere som ønsker et verktøy for mer nøyaktige vurderinger og rådgivning til kundene sine. De forventer at tjenesten gir en konkurransefordel i markedet.

### Tentativ Tidslinje med Milepæler

Prosjektet følger en komprimert tidslinje med følgende milepæler:

- **Milepæl 1:** Prosjektoppstart og datainnsamling
  - Definere mål, rollefordeling og samle inn nødvendige data om boligpriser og markedstrender.
- **Milepæl 2:** Modellutvikling
  - Utvikle en grunnmodell for boligprisestimering basert på tilgjengelig data og sikre modellens nøyaktighet.
- **Milepæl 3:** Nettsideutvikling og integrasjon
  - Bygge en enkel nettside for innsamling av boligdata fra brukerne og integrere modellen for generering av estimer.
- **Milepæl 4:** Testing, rapportskrivning og overlevering
  - Utføre testing, dokumentere resultater og levere prosjektet med rapport på Canvas.

### Ressurser

- **Personell:**
  - **Utvikler:** Samler data og utvikler modellen for boligprisestimering.
- **Beregningsressurser:**
  - **Kraftig laptop eller arbeidsstasjon:** Til lokal utvikling og testing.
  - **Skyplattform for maskinlæring:** Til modellkjøring hvis databehandlingen krever ekstra ressurser.
- **Dataressurser:**
  - **Bolig- og markedsdata:** Historiske data om boligpriser og trender for modelltrening og nøyaktige estimer.

### Data

Prosjektet baserer seg på et treningssett og et testsett som inneholder de mest nødvendige variablene for nøyaktige boligprisestimer. Det er ingen planer om å oppdatere datasettene regelmessig, da prosjektets behov anses som dekket med de opprinnelige dataene. Ytterligere datakilder, som Matrikkelen og Grunnboken, kunne økt påliteligheten, men er utelatt for å forenkle prosjektet og unngå personvernproblematikk.

For å sikre konsistens i dataene, kan kryssvalidering med alternative kilder brukes. Prosjektet samler ikke inn personopplysninger, og det anses derfor å oppfylle personvern hensyn. For å forbedre modellens ytelse, vurderes tillegg som min-max-skalerting for numeriske variabler og beregning av pris per kvadratmeter for å fange opp verdiforskjeller.

## Modellering

Random Forest Regressor er valgt som hovedmodell for boligprisestimering grunnet god ytelse og tolkningsmuligheter med minimal parameterjustering. Alternativt vurderes Gradient Boosting og XGBoost, som begge kan bidra med spesifikke fordeler: Gradient Boosting er effektivt og enkel å implementere, mens XGBoost gir høyere nøyaktighet, men krever mer tuning.

Modellen evalueres ved hjelp av MAE, MSE, RMSE og  $R^2$ . For at modellen skal anses som klar for produksjon, bør MAE være innenfor 10 % av gjennomsnittlig boligpris, og  $R^2$  bør være 0,80 eller høyere. Feature importance analyseres med Random Forest og eventuelt XGBoost for å identifisere hvilke variabler som har størst innvirkning på prediksjonene.

## Deployment

Distribusjonen av modellen vil gjøres gjennom Gradio, som gir en enkel brukergrensesnittløsning for boligprisestimerer. For å sikre en brukervennlig opplevelse, settes et mål om å holde responstiden under fem minutter, slik at brukerne raskt får prediksjonene sine.

Modellen vil være åpen og tilgjengelig uten tilgangsbegrensninger, og ingen prediksjonsdata lagres, noe som forenkler sikkerhetskrav og personvern. Den minimale overvåkingen som kreves betyr også at det ikke er behov for retrening eller regelmessig vedlikehold av modellen.

## Referanser

Dataene er hentet fra GitHub, og verktøy som Google Colab, Kaggle, ChatGPT og Google Drive har vært nyttige i datahåndtering og utvikling. Valg av Random Forest-modellen er basert på pensum i faget, da den gir et pålitelig verktøy for prediksjon av boligpriser i denne akademiske konteksten. Juridiske krav som GDPR er ikke relevante, ettersom ingen personopplysninger håndteres i prosjektet.