Estimate the value of a used car

Emil Tveten, Anders Mikkelsen, Vlad Craiu. 02.11.2025

1: BESKRIV PROBLEMET

OMFANG / SCOPE

Målet er å bygge en modell som predikerer salgspris for brukte biler basert på opplysninger som merke/modell, års-modell, kjørelengde, drivstoff og tilstand. Prediksjonene kan støtte prisvurdering for selgere/kjøpere (anbefalt pris) og forhandlere (innkjøp/prising), samt brukes i portaler for automatisk prisestimat.
Hvorfor ML? Uten maskinlæring settes pris ofte ved manuell sammenligning med lignende annonser, eller tidligere salg av lignende solgte biler. Dette er tidskrevende, kan endre seg årlig og er inkonsistent. En datadrevet modell kan lære ikke-lineære sammenhenger (f.eks interaksjoner mellom årsmodell og kilometerstand, driftstofftype, girtype).

Bruksscenarioer:

- Nettportal viser “estimert pris” når en bruker legger inn bilinfo.

METRIKKER

- **ML-metrikker:** RMSE, MAE, R^2 på hold-out validering og kryssvalidering.

2: DATA

Kilde og Datatyper

- **Dataset:** Datasett kom fra Kaggle [**Regression of Used Car Prices**].

- **Størrelse:**
    - Train: 188 533 rader, 13 kolonner.
    - Test: 125 690 rader, 12 kolonner (uten “price”).

    - **Kategoriske:** Brand, model, fuel_type, engine, transmission, ext_col, int_col, accident, cleant_title.
    - **Numeriske:** model_year, milage, (Id).

    - **Manglende verdier:** Totalt 28 954 manglende felter. Mest i “clean_title” (21 419), deretter “fuel_type” (5 083) og accident (2 452).

    - **Målvariabel (price):** n = 188 533, median = 30 825, mean = 43 878, min/max = 2 000 / 2 954 083.

Databehandling og Rensing

- **Rensing:** fjerning av data som ikke gir prediktive verdier (Id, engine, int- og ext_col).

- **Manglende verdier:**
    - **Numbersike:** median-imputering.
    - **Kategoriske:** manglende verdier får egen kategori “Unknown”.

- **Skalering/enkoding:**
    - One-hot/target-encoding for høykardinalutet (modellvariant).

- **Avvik/outliers:** pris har en lang høyrehare, sammenlignet også log-transform som alternativ treningsmål.

- **Train/valid/test:** Splittet i test og valid. Det ble brukt [80/20].

3: MODELLERING

Kanditatmodeller

1. Linear regression: Antar lineart forhold mellom features og pris. Rask og tolkbar, men sårbar for uteliggere og fanger ikke ikke-linearitet.

2. Ridge Regression: Lignende på Linear regression, men med L2-straff som krymper koeffisienter og reduserer varians ved mange features.

3. Random Forest: et samspill av beslutningstrær trent på bootstrap-utvalg og feature subsampling, Til ikke-linearitet og interaksjoner, robust mot støy, lite tuning-sensitiv.

4. Model Tree Regressor: Decision tree hvor hvert blad har en lokal lineær model (danger piece-wise linear struktur). Mer tolkbar enn “svarte bokser” og kan gi lavere feil enn rene trær når forhold er lokalt lineære.

5. XGBoost (valgt): Gradient-boosting av svake trær. Lærer seg sekvensielt på rest og gir ofte best ytelse på tabulære data. Har innebygget regularisering og støtte for ubalanse/vekter.

Endelig valg av modell (XGBoost)

- **Hyperparameter:** n_estimators=400, learning_rate=0.05, max_depth=8, subsample=0.8, colsample_bytree=0.8, random_state=42, tree_methos=”hist”, n_jobs=-1.

Resultat

Modell	MAE	RMSE	R^2
XGBoost (price)	17 278.10	68 782.53	0.149
XGBoost(log-skala)	0.3476	0.4978	0.654

4: DEPLOYMENT

Streamlit

Modellen ble gjort tilgjengelig med en webapp laget i et .venv i pycharm med Streamlit. Appen lar brukerne velge bilmerke, modell, årsmodell, hvor langt den har kjørt, drivstofftype og volum på motor. Webbappen predikerer da prisen.
Selve modellen er en XGBBoost-pipline som lastes inn ved oppstart, i tillegg hentes hjelpe data som liste over modellene til hvert bilmerke.

Prediksjonen skjer i det brukeren trykker på “Estimer pris”, resultatet vises i NOK og USD.
Appen kan kjøres lokalt eller på https://dat158gr34.streamlit.app/.

For videre arbeid kan det utvides til at man logger prediksjonene, automatisk oppdatering av modellen, når/om nye data blir tilgjengelig.
Man kunne også laget et API, slik at andre tjenester kan bruke det.

5: REFERANSER

Kaggle dataset: Regression of Used Car Prices
Notebooks fra https://github.com/HVL-ML/DAT158