USED CAR PREDICTOR

Ola Haveland Brenne, Ridwan Saalah Mohammed Ali

1: BESKRIV PROBLEM

Omfang

Formålet med prosjektet er å lage ei web-basert løysing som kan estimere marknadsprisen til brukte bilar ved hjelp av maskinlæring. Modellen tar inn informasjon om bilens merke, modell, årsmodell, kilometerstand, motor, drivstofftype, girtype, eksteriør- og interiørfarge, uhellshistorikk og "title"-status, og returnerer ein estimert verdi.

Dette er nyttig fordi prisar for brukte bilar varierer kraftig basert på fleire faktorar, og kan vere vanskeleg å vurdere manuelt for både kjøparar og seljarar. Maskinlæring er ein lovande metode fordi slike prisrelasjonar er komplekse og ikkje lett beskrivne med enkle reglar.

I dag blir problemet ofte løyst ved å sjå på prisar på Finn.no, Autotrader, bruktbilforhandlarar, eller å bruke eigne erfaringar. Dette er tidkrevjande og subjektivt. Ei maskinlæringsløysing kan gi raskare og meir konsistente estimat.

Produktet kan oppnå marknadsverdi ved å spare tid, gi meir rettferdig prissetting, og redusere risiko for feilprising. Moglege brukarar inkluderer privatpersonar, bilforhandlarar og forsikringsselskap.

For å gjennomføre prosjektet trengst ein utviklar, Python-miljø for modelltrening, FastAPI for server-delen, og enkel hosting av webgrensesnitt.

Metrikker

For å måle kvaliteten på modellen brukte eg:

- RMSE (Root Mean Squared Error) — måler typisk avvik mellom predikert pris og faktisk pris.
- MAE (Mean Absolute Error) — viser gjennomsnittleg absolutt feil.
- R^2 Score — viser kor mykje variasjon modellen forklarer.

Desse gir eit klart, numerisk mål på korleis modellen presterer. For å vere nyttig i praksis bør feilen vere låg nok til at estimatet er relevant (f.eks. < 15–25% feilmargin av typisk bruktbilpris).

2: DATA

Datasettet kjem frå Kaggle Playground Series – Season 4, Episode 9, som inneheld reelle bruktbiloppføringar med tilhøyrande salsprisar. Dataen består av tabulære, strukturerte finessar med både numeriske og kategoriske kolonnar.

Dette er eit supervised learning-problem, der målet (label) er salgspris. Labelane er allereie inkludert i datasettet, noko som gjer dette effektivt å trene på. Ein kan anta at labelane representerer reelle, historiske prisdata.

Datasettet krev litt datareinhald:

- handtering av manglande verdiar
- konvertering av kategoriske felt til one-hot encoding
- standardisering av numeriske verdiar
- fjerne ID-kolonnar som ikkje bidrar

Etiske vurderingar:

- ingen persondata eller sensitive identifikatorar er inkludert
- datasetet er offentleg og lovleg distribuert

Dataen er representert i ein pandas DataFrame og behandla via scikit-learn sin ColumnTransformer.

3: MODELLERING

Eg utforska modellar som passar tabulær data, inkludert:

- Linear Regression (baseline)
- Random Forest Regressor
- HistGradientBoostingRegressor (valgt modell)

For baseline samanlikna eg med ein enkel modell som returnerer gjennomsnittsprisen. Deretter vurderte eg meir avanserte modellar for betre ytelse.

HistGradientBoostingRegressor presterte best, og toler kategoriske data etter preprocessing. Feature importance viser at model_year, milage, brand og model dominerer prisinformasjonen. Andre finessar, som clean_title og accident, har svakare effekt i dette datasettet.

For model tuning og validering brukte eg ein tren/test-splitt (~80/20). Modellens ytelse vart analysert via RMSE, MAE og R^2. I tillegg undersøkte eg prediksjonsfeil og evaluerte korleis ukjente kategoriverdiar kan føre til null-innflytelse via handle_unknown='ignore'.

4: DEPLOYMENT

Modellen er eksportert som .pkl ved hjelp av joblib og lastast inn av ein FastAPI-server. API-et tilbyr:

- /predict — tar inn bilfinessar og returnerer pris
- /schema — returnerer finessestruktur
- valfri debug-modus for feilsporing

Frontenden er ei enkel webside med HTML/CSS/JavaScript. Brukaren fyller inn bilinformasjon og får eit prisestimat. Pris kan konverterast mellom USD og NOK gjennom ein eigen valutakalkulator.

Modellen kan enkelt vedlikehaldast ved periodisk retrening med ferskare data. Over tid kan nye finessar leggast til (t.d. horsepower eller marknadsregion).

For vidare forbetring:

- hente sanntidsdata frå API-kjelder
- betre feature-engineering (uttrekke liter/hestekrefter)
- meir robust normalisering av kategoriske felt
- lage dropdown-menyar frå treningsvokabular

5: REFERANSER

- Kaggle Playground Series S4E9 Used Car Prices Dataset
- Scikit-learn documentation
- FastAPI documentation
- Tailwind documentation (CSS styling)
- Kaggle community notebooks/inspirerende eksempel
- Lecture material from DAT158