Prispredikering for bilmarkedet.

Erlend Hollen Assen & Christian Gunvaldsen, 14/11/2024

Prosjekt beskrivelse

Prosjektet har som mål å utvikle et verktøy som kan forutsi salgsprisen på brukte biler på markedet. Ved å legge inn spesifikke bilopplysninger som merke, modell, årsmodell, kilometerstand, drivstofftype, girtype og andre relevante egenskaper, skal verktøyet hjelpe selskapet med å optimalisere profitt på videresalg av kundebiler.

SCOPE

Oversikt

- Utnytte offentlig salgsdata og interne bedriftsdata for å bygge et maskinlæringsverktøy som kan estimere bruktbilpriser.
- Strukturere data slik at det er mulig å kategorisere den konsekvent, noe som hjelper modellen med å gi nøyaktige prediksjoner.

Brukstilfelle

Verktøyet vil bli brukt av selskapet for å sette konkurransedyktige og lønnsomme priser på brukte biler i sitt lager, noe som forbedrer beslutningsprosesser.

METRIKKER

Pris antydning

Vår løsning skal ha et maks kvadratisk avvik på $RMSE = (\frac{1}{N}\sum_{i=1}^{N}(yi - \widehat{yi})^2)^{\frac{1}{2}}$

Forretningsmetrikker:

- Sikre at prisestimatene er nøyaktige innenfor en 10 % feilmargin for 90 % av tilfellene og presis treff sikker.

Maskinlæringsmetrikker:

- RMSE (root mean squared error) som en indikator på modellens ytelse og prediksjonsfeil med $RMSE = (\frac{1}{N}\sum_{i=1}^{N}(yi - \widehat{yi})^2)^{\frac{1}{2}}$
- Måle modellens latens og gjennomstrømming, selv om disse er sekundære metrikker ved lokal bruk.

DATA

Data Oversikt

Vi har data fra en stor datasamling av offentlig og bedriftens egne solgte bilpriser.
Vår data består av 188.533 biler solgt på bruktbilmarkedet av godt varierende modeller og stand, og hva bilene ble solgt for, i vår data har vi: Bilmerke, Modell, årsmodell, milstand, drivstofftype, motortype, girkassetype, eksteriørfarge, interiørfarge, om bil har vært registrert kollisjon, og om bilen har en ren tittel, pris solgt på markedet.

Rensking av data

- Manglende verdier ble håndtert ved å erstatte dem med relevante plassholdere (f.eks. "Ukjent" for manglende titler).
- Kategoriske variabler ble numerisk kodet for kompatibilitet med modellen.
- Laget nye verdier felter som vi mener manglet for å komme til et optimalt svar som alder og kilometer per år.

MODELLERING

Valg av Modell

Med tanke på hvilke dataer og hvordan vår data kommer til å bli brukt, vil det være naturlig at vi velger et algoritme som vekter alle verdiene etter som det er en sammenheng mellom alle. og hvis vi har mye verdier kategoriserbare verdier vil det være naturlig å prøve “RandomForestRegressor”, “GradientBoostingRegressor” samt prøve “XGBRegressor”.
I vårt prosjekt har vi valgt å gå for å kjøre et “RandomForestregression algoritme” siden det var den vi fikk best resultat på prediksjonen når vi kjørte første utkast og valgte å videreutvikle denne.

Pipeline

Vi valgte å sette opp en pipeline for å gjøre at når vi bruker modellene våres at det er lettere å kjøre disse etter som med en pipeline vil vil kunne sette opp at vi strukturerer dataen slik vi ønsker ønsker og vil slippe å prosessere dataen på hvert av stegene, siden vi lager en jobb som vi kaller som preprosseserer dataen vi får inn på samme måte som vi formaterer treningsdataen.

- En preprocessing-pipeline ble satt opp for å sikre konsekvent datahåndtering under trening og testing.
- Dette gjør det mulig å kjøre forhåndsbehandling og modellprediksjoner sømløst, reduserer redundans i koden og forbedrer modellvedlikehold.

Implementering

Modellen ble trent og evaluert i en lokal Python Jupyter Notebook. Siden prosjektet er utforskende, ble det ikke gjennomført full implementering. Fremtidige steg kan innebære å distribuere modellen som en mikrotjeneste dersom selskapet ønsker en integrert web-basert løsning.

DEPLOYMENT

Vi valgte å kjøre våres kode gjennom Python Jupyter notebook og kjøre dette lokalt på datamaskinen vår, ettersom dette var lettere og vi ikke så et behov for å sette det opp på en faktisk reell nettside, siden det ikke er noe vi planlegger å vedlikeholde.

REFERANSER

Your mom
https://docs.jupyter.org/en/latest/
https://scikit-learn.org/stable/user_guide.html
https://chatgpt.com/
https://docs.python.org/3/library/
https://www.kaggle.com/competitions/playground-series-s4e9