Sqick sqack på tide å selge bilen?

Sindre Bakke Marthinussen, 15.11.2024

BESKRIV PROBLEMET

SCOPE

Målet med prosjektet jeg satt meg var å lage en første iterasjon av en maskinlæringsmodell for å kunne gi en pekepinn på pris. Dette skal den gjøre igjennom visse metrikker som merke, årsmodell, farge (inne og ute), ulykke, dokumentasjon og salgspris.
Vel dette produktet kan man i første runde hjelpe selgere fra amatør som selger på finn til en proff bilselger. Det å kunne få raske pekepinner på pris kan hjelpe med å få økt effektiviteten på pris giving til en brukt bil. En versjon kan også lages til å være en individuell nettside og/eller innlemmes i en side som finn.no.
Hvis man utvikler et eget produkt som app og egen nettside så har man få stakeholders i prosjektet annet enn seg selv, de ansatte og eventuelle inversorer.
Og for å få investorer til dette så må man ha en tentativ plan på hvordan man skal jobbe seg framover mot milepæler. Vel første milepæl er å ha et minimum produkt for testing internt og som prototype til mulige kunder. Den sekundere større milepælen vil bli å kunne «publisere» det første enkleste brukbare produkt enten om det vil være en nettside og/eller en app.

METRIKKER

Vel en veldig enkel måte å teste dette prosjektet her for sin ytelse er å kunne bli med i konkurransen som dataene i dette prosjektet er hentet ut ifra. Med å teste opp mot dette så er det lettere å kunne sette og vite når man har nådd minimumskravet. På prosjektet nå så ble kvadratiskavvik for å estimere forskjellene mellom faktisk verdi og estimert.

DATA

Dataene som ble brukt til å estimere bruktbil pris er hentet fra Kaggle. Disse dataene inneholder informasjon som modell, modellår, drivstoff, motor, girkasse, eksteriør farge, interiør farge, ulykke, ren tittel (engelsk clean title) og pris. Det var også en id rekke som ble raskt droppet.

Dataen er tabellformet som et Excel ark, men det var ikke like uniformt som et Excel ark i datatypen. Så det var en god blanding av datatyper som Integer og String.
Men hva skal en forvente av data som er hentet ut fra en offentlig nettside som Kaggle «Regression of Used Car Prices». Mengden data som er hentet ut fra den nettsiden består av hele 125690(test)+15907(train).
Den mengden data er tilstrekkelig for å trene en første iterasjon av produktet.
Og selvfølgelig kunne det ha hjulpet å ha mer detaljert informasjon og flere parametere å velge ifra. Dataene som ble ut ifra dette måtte det gjøres en kvalitetskontroll på, dette var for å sørge for at de var konsistente nok. Ettersom at jeg støtte på «pipeline» problemer ble det funnet ut at det måtte gjøres datakonverteringer med label encoding. Naturligvis ble det også gjort ekstra tid på datakvalitetskontroll for å korrigere feil i datasettet.

MODELLERING

På prosjektet ble det testet ulike typer modeller som Random Forest Regressor, Gradient Boosting, Linear Regression, XGBClassifier som ledet meg til XGBoost Regressor også var det noen flere som ikke har blitt notert ned.

DEPLOYMENT

I dette prosjektet har det blitt brukt gradio for deployment. I dette kan en motta brukerinput og bruker kan bruke modellen for å få et prisestimat.
Videre planer for overvåkning er akkurat nå ikke eksisterende siden prosjektet regnes som «deprecated» fra min side, videre utvikling ville jeg sikkert begynt på en versjon 2 med mer erfaring.

REFERANSER

Inspirasjon for løsning hentet og andre jeg som ikke huskes:
https://www.kaggle.com/code/martinapreusse/ps4e9-cat-svr-lgbm-nn-py
«for retting av skrivefeil, bugs, inspirasjon»
https://chatgpt.com/