Used Car Prices

Eric Da Costa Kobbeltvedt & Johan Severin Reitan / ML project 2, 18.10.2024

BESKRIV PROBLEMET

SCOPE

Målet med prosjektet er å utvikle en maskinlæringsmodell som kan forutsi bruktbilpriser basert på ulike bilattributter som merke, modell, årsmodell, kilometerstand, bensintype, motor, girkasse, utvendig farge, innvendig farge, registrert ulykke, dokumentasjon og salgspris.

Dette prosjektet skal hjelpe til med å sette mer nøyaktige og konkurransedyktige priser for brukte biler. Det gjør denne prosessen i dag er ofte en tidkrevende og vanskelig prosess. Ved å ta i bruk en slik modell kan vi øke eventuelle salg, og redusere tiden som blir brukt på prising av brukte biler som igjen vil føre til økte inntekter. De som kan dra nytte av dette er bruktbilforhandlere og privatpersoner som ønsker å selge bruktbiler.

Prosjektet vårt er designet som en frittstående løsning og er ikke en del av et større system eller pipeline. Maskinlæringsmodellen vår vil bli brukt direkte av brukere gjennom et enkelt grensesnitt. Selv om prosjektet ikke er en del av et større system, består løsningen av følgende hovedkomponenter: dataforbredelse, modelltrening og brukergrensesnitt.

Dataforbredelse: Rensing og konventering av data for å gjøre dem klar for modelltrening. Dette vil si fjerning av manglende verdier og konventering av kategoriske variabler til numeriske verdier.

Modelltrening: Trening av maskinlæringsmodellen ved hjelp av de behandlete dataene vi har. Vi bruker slagprisene for de ulike bilene som lables for å lære modellen til å forutsi priser basert på bilens egenskaper.

Brukergrensesnitt: Et enkelt grensesnitt hvor brukeren kan legge til de ulike egenskapene til bilen og få en prisvurdering fra modellen.

METRIKKER

Vi tok utgangspunkt i topplisten på Kaggle for å se hvor vi lå i forhold til andre gode løsninger. Kvadratisk avvik ble brukt for å finne forskjellene mellom faktisk verdi og estimert verdi.

DATA

Data som vi har valgt å ta i bruk er for å kunne estimere en brukt bilpris. Data inneholder informasjon om selve bilen som merke, modell, års modell, kilometerstand, bensintype, motor, girkasse, utvendig farge, innvendig farge, registrert ulykke, dokumentasjon og slagspriser.

Dataen er tabellformede data med rader og kolonner som inneholder en blanding av datatyper som “String” og “Integer”. Dataene er hentet fra en offentlig tilgjengelig dataset på Kaggle som heter “Regression of Used Car Prices”. For øyeblikket har vi til gang på 188.533 data punkter og 13 funksjoner (kolonner) som vi forventer skal være tilstrekkelig for å kunne trene en modell.

Ved at vi tar i bruk et datasett som allerede inneholder salgspriser som labels, sikrer vi at modellen trenes på reelle og relevante data. For å sikre at labels er tilstrekkelig konsistente kan vi gjøre ulike tiltak som datakvalitetskontroll, flere datakilder og statiske metoder. Vi har i stor grad gjort konventeringer av kategoriske variabler fra strengeverdier til numeriske verdier ved hjelp av label encoding for å sikre at dataene er i en form som kan brukes effektivt av maskinlæringsmodellen. Brukt også datakvalitetskontroll som er å ha en grundig gjennomgang av datasettet for å finne og korrigere eventuelle feil eller inkonsekvenser i salgsprisene.

Vi forsøkte å gjøre modellen best for “folk flest” og luket ut utliggere i pris og årsmodell. Slik fikk vi litt mindre data å trene modellen med, men et bedre gjennomsnittsresultat.

I tillegg samlet vi sammen liknende verdier i kategoriene; transmission, fuel_type, ext_col og int_col. Dette ga oss litt bedre oversikt over hva de forskjellige verdiene faktisk påvirket prisen.

MODELLERING

Vi testet flere ulike modeller, inkludert Linear Regression, Random Forest Regressor, Gradient Boosting Regressor og XGBoost Regressor. Etter vi evaluerte resultatet til de ulike modellene valgte vi å benytte XGBoost Regressor, da denne modellen ga det beste resultatet med en RMSE-verdi (Root Mean Square Error) på 26 613.

DEPLOYMENT

Ved bruk av Gradio kan en bruker taste inn spesifikasjonene til bilen sin og få et prisestimat fra modellen vår.

REFERANSER

uthenting av motor-specs via regulære uttrykk:
https://www.kaggle.com/code/dvaled/used-car-regression-stack-regressor