Rapport House price prediction

Første jeg vil gjøre, og noe som vi alltid må gjøre innen maskinlæring, er å utforske datasettene vi har fått utdelt. Jeg vil da først og fremst gå gjennom både testsettet og treningssettet. Da vil jeg først se hvilke kolonner vi har og vurdere hvor viktige de er i forhold til modellen vår. Man må ofte tenke litt større og vurdere situasjonen denne modellen skal trenes opp mot. Man må tenke realistisk og spørre seg selv hvorfor de ulike attributtene er viktige å ha med eller ikke.

Så er det å rydde i dataene våre. Det innebærer å gå gjennom verdiene for de ulike attributtene og vurdere om de er realistiske. Deretter må vi få en oversikt over nullverdier. Noen attributter kan ofte inneholde mange nullverdier, så vi må vurdere om vi skal fjerne dem.

Jeg begynte med å rydde opp i treningsdataene. Da begynte jeg med å fjerne alle nullverdier. Kolonnene som hadde mer enn 50% nullverdier valgte jeg å fjerne, da dette ville påvirke resultatet. Jeg erstattet alle nullverdier som enten var flyttall eller heltall med gjennomsnittet av alle tallene i den kolonnen. Så erstattet jeg alle de kategoriske verdiene med den hyppigste typen av den kolonnen. Når jeg hadde blitt kvitt de fleste nullverdiene, fjernet jeg alle rader som inneholdt nullverdier for å være sikker på at jeg ikke hadde noen igjen. Alt dette gjorde jeg også med testsettet, men jeg fjernet ingen rader siden jeg må ha like mange rader som i sample submission-filen.

Så måtte jeg gjøre alle de kategoriske kolonnene om til dummykolonner, altså kolonner som inneholder tall i stedet for kategorier. Dette er fordi maskinlæringsmodeller håndterer tall mye bedre enn bokstaver osv. For at trenings- og testsettet skulle ha nøyaktig like dummyverdier, gjorde jeg en concat, slik at de delte antall kolonner. Så sendte jeg dem gjennom en funksjon som omgjorde kategoriene til tall. Så splittet jeg dem opp igjen.

Videre skulle jeg velge en algoritme og gjøre en prediksjon. Jeg endte da opp med XGBoost-algoritmen, eller gradient boosting som det heter etter litt research. Datasettet mitt var ganske komplekst, noe XGBoost er kjent for å håndtere.

For å optimalisere algoritmen gjorde jeg en hyperparameter tuning. Dette går ut på å finne de hyperparameterne som passer best til situasjonen. Da kan vi finne dypden i algoritmen, antall noder, learning rate osv. Dette vil som sagt optimalisere modellen min slik at vi kan få en bedre prediksjon. Så var det å hente ut den beste modellen og trene den med X-variablene og y-variabelen. Når den var trent, kunne jeg plotte inn testdata og få ut en liste med prediksjoner. Gøy!

Så kombinerte jeg bare Id-kolonnen fra sample submission-filen og testdata og lagret dette i en CSV-fil. Dette leverte jeg på Kaggle og fikk en Root-Mean-Squared-Error score på 0.13103, noe jeg er veldig fornøyd med.

Til slutt laget jeg en Gradio-side der du kan legge inn en fil for et hus og så får du ut prediksjonen på hva huset kan gå for. Ganske kult, vil jeg si.

Refleksjon: Gjennom dette prosjektet har jeg lært viktigheten av å utforske og rydde i datasettene du får utdelt. Det er svært viktig at modellen blir trent opp riktig og at med bare små justeringer kan man få et mye bedre resultat. Det har vært ganske utfordrende, men lærerikt.