Dataset: Medical Cost

Regression

Markus Kinn

I dette dokumentet kommer jeg til å diskutere valg jeg har tatt og sammenligne resultater fra ulike algoritmer. Jeg har ett dokument for hvert datasett.

For å forbedre resultatene på modellene har jeg prøvd å implementere så mye so mulig fra *Machine Learning Performance Improvement Cheat Sheet.*

**Generelt om datasettet:**

Datasettet går ut på å predikere medisinske kostander for en person ut ifra disse attributtene: age, sex, bmi, children, smoker og region.

Datasettet har ingen feil jeg trenger å rette opp i.

**XGBoost:**

For dette datasettet var det ikke behov for å fjerne noen kolonner eller gjøre noen annen form for data manipulasjon før jeg trente modellen med XGBoost.

**Før tuning:**

**Text

Description automatically generated**

**Valg av tuning strategi:**

For XGBoost valgte jeg nok en gang å teste ut 3 tuning strategier: sekvensiell, mindre sekvensiell og random search

(tar kun den beste scoren gjennom mange gjennomganger for Random search).

**Etter tuning (Sekvensiell):**

Text

Description automatically generated

**Etter tuning (Mindre sekvensiell):**

Text

Description automatically generated

**Etter tuning (Random search):**

**Text

Description automatically generated**

**Diskusjon av resultater:**

Baseline modellen ble ekstremt over-fitted, men generelt gode resultater for det.

Etter tuning, med alle strategiene, klarte jeg å nærmest eliminere over-fittingen, samtidig som jeg økte cross val, R2 og scoren på trening og test data. Redusering av over-fitting kom mest sannsynlig som et resultat av redusert max\_depth fra baseline modellen.

For dette datasettet presterte den sekvensielle strategien best på alle scorene jeg tok med. Dette kommer mest sannsynlig som et resultat av at min mindre sekvensielle metode ble testet for dårlig, samt at Random search fikk tilfeldigvis ikke bedre score.

**Random Forest:**

Gjorde ikke noe ekstra med datasettet her heller.

**Før tuning:**

**Graphical user interface, text

Description automatically generated**

**Valg av tuning strategi:**

Her prøvde jeg både min egne litt mindre sekvensielle metode og Random search. Opprinnelig fikk jeg EKSTREMT dårlig resultat med Random Search, men etter å ha fjernet ‘max\_leaf\_nodes’ og ‘max\_samples’ fra Griden så økte resultatet til å bli tilnærmet lik de andre.

**Etter tuning (Mindre sekvensiell):**

**Text

Description automatically generated**

**Etter tuning (Random Search):**

**Text

Description automatically generated**

**Diskusjon av resultater:**

En ekstremt over-fitted baseline modell, men med noe høyere test score og litt lavere training score enn hva XGBoost baseline modellen fikk.

Her var det min egne mindre sekvensielle tuning strategi som presterte best. Etter tuning klarte jeg å nærmest eliminere over-fitting ved å øke test score og senke training. Dette kom mest sannsynlig som et resultat av lavere max\_depth enn hva baseline modellen hadde.

**Support Vector Machine:**

Jeg måtte skalere dataen for denne algoritmen. For dette brukte jeg RobustScaler() funksjonen. Grunnen til at dette er viktig for SVM er at den bruker distansen mellom observasjoner til å predikere (veldig kort fortalt).

**Før tuning:**

**Text

Description automatically generated**

**Valg av tuning strategi:**

Her brukte jeg kun et vanlig grid search da algoritmen ikke har mange parametere å tune

**Etter tuning:**

**Text

Description automatically generated**

**Diskusjon av resultater:**

Baseline modellen fikk ekstremt dårlige resultater, noe også tuning modellen fikk. Etter litt undersøkelser regner jeg med at dette kommer fra dårlig fordeling av noe av de sentrale variablene i datasettet, som for eksempel smoking.

**Sammenligning av alle resultater:**

For dette datasettet klarte XGBoost og Random Forest seg ganske bra. Begge modellene fikk akseptable mengder med over-fitting etter tuning, samt generelt bra scorer på de måleenhetene jeg tok med.

RF tok seieren på dette datasettet på alle scorene. Spesielt viktig på regresjons problemer er R2 scoren, som måler den absolutte gjennomsnittlige residualen.