# Maskinlæring Prosjekt 2
# Tore Tveita, 18/11/2022

# BESKRIV PROBLEMET
# SCOPE

Prosjektets mål er å gjennomgå et maskinlæring prosjekt hvor man starter fra en problemstilling og et datasett, produserer og trener en modell for problemet, og til slutt utplasserer den utviklede modellen. Selve problemet modellen skal prøve å løse er nokså valgfritt. Jeg har valgt å prøve å lage en modell i forbindelse med en konkurranse på nettsiden Kaggle. Problemet som i den konkurransen skulle løses best mulig er å lage en modell som kan gjennom data fra filmdatabasen TMDB forutsi hvor mye filmen kommer til å ha i inntekt. Data inkluderer budsjett, språk, popularitet, sjanger og mer. Sluttmålet med dette prosjektet når det er ferdigutviklet er å ha en fungerende modell som kan med relativt høy presisjon fortelle om en film, gitt visse attributter, kommer til å ha økonomisk suksess eller ikke. En slik modell må kunne være pålitelig, og vil måtte kunne håndtere store mengder forskjellige datatyper. Interessenter for en slik modell om den er veldig vellykket vil være filmindustrien selv, som kan bruke den til å unngå risiko for økonomiske tap. Det vil være behov for personell både for utvikling og vedlikehold.

# METRIKKER

Det viktigste for at prosjektet skal kunne bli brukt på en suksessfull måte er at de som bruker det kan stole på at resultatet fra modellen er nokså nøyaktig. Den bør ha høy treffsikkerhet i alle tilfeller, dette er viktigere enn effektiviteten til modellen. Bommer modellen kraftig kan dette ha store økonomiske virkninger for brukeren, men om modellen bruker lang tid har mye mindre konsekvens.

# DATA

Dataen modellen skal kunne håndtere flere datatyper. Mye av dataen som tilhører filmer er kategoriell, ting som sjanger, språk, regissør og skuespillere. Mye av dette kan være like viktig som numerisk data som budsjett, popularitet og spilletid. Modellen må kunne bruke mye eldre data, ettersom det er begrenset mengde “fersk” data på film. Veldig mye av denne dataen er publisert på internett og kan enkelt inkluderes i trenings dataen. Det som er spesielt viktig i forhold til dette problemet er hvordan data labeling vil være. SIden inntekter til en film ikke umiddelbart vil være synlig, det kan ta år før den slutter å vokse, er det viktig at modellen på et vis tar dette i betraktning. Derfor er det viktig å bruke både mye eldre data for dette, altså unngå helt nye filmer som kan påvirke modellens evne til å evaluere om den har gjort riktig i forutsetningen. Derfor er det viktig at modellen heller ikke trenes på data fra helt nye filmer. Dette er noe som må analyseres nærmere for å finne best mulig løsning på.

# Personvernhensyn

I forhold til personvern er det viktig å påpeke at dataen modellen skal bruke er kun publisert informasjon om filmer og personell. .

# Hvordan skal data representeres for maskinlæringsmodellene?

Som nevnt vil modellen møte flere problemer med manglende data. Det er derfor viktig at den skal kunne bruke de tilgjengelige dataene på best mulig måte, og ikke regne med data som kan skade prediksjonen. For å hjelpe med dette kan også kombinasjoner av data og data forhold gi mer informasjon enn bare de numeriske verdiene. Det er også et behov for å håndtere store mengder kategorisk data, som må omgjøres til et numerisk format. Et eksempel på dette kan være gjennom one-hot encoding av disse dataene.

# MODELLERING

Den endelige modellen jeg nå har utvikler bruker grid search for å finne beste kombinasjoner av hyper parametre. Grid search funksjonen som er implementert bruker random forest regresjon som er den modellen som gav best resultater av de jeg prøvde. Gjennom flere iterasjoner av dette har bedre og bedre parametre blitt funnet som har gitt bedre og bedre resultater i den endelige prediksjonen. I forhold til videreutvikling av dette burde flere modeller utforskes, med flere mulige parametere og mer “fine-tuning”. Det er også behov for

en metode å oppdage feil-prediksjoner med å sammenligne prediksjoner med faktiske data slik at man kan forstå hvor modellen kan forbedre seg. Det er også viktig å implementere en metode for å evaluere modellen, slik at det kan oppdages om den gir dårligere resultater ettersom den får mer treningsdata.

# DEPLOYMENT

Jeg forsøkte selv å sette modellen i drift gjennom en lokal flask webapplikasjon, men fikk ikke tid til å få den til å fungere. Likevel var dette tanken at prosjektet skulle resultere i. Videre på dette skulle det implementeres mulighet for å sende inn en films attributter i et skjema og få tilbake svar fra modellen om forventet inntekt. Dette skulle være en enklere fremstilling av hvordan det ferdigstilte prosjektet skulle driftes, at brukere skal kunne sende inn data, for eksempel en filmtittel og få tilbake svar. Dette hadde krevd at nettsiden også hadde tilgang til en database av filmer, f.eks imdb. Dette er også noe som hadde hindret noen i å sende inn data som ikke er realistisk, og eventuelt kan skade modellens treningsevne. Dette hadde likevel krevd vedlikehold av treningsdata, og krevd en menneskelig faktor til dette. Samtidig må en server kjøres for å holde nettsiden tilgjengelig for alle, enten gjennom investering av hardware eller en skytjeneste for å hoste webapplikasjonen.

# REFERANSER


Géron, A. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. O'Reilly.

*TMDB Box Office Prediction*. (n.d.). Kaggle. Retrieved November 18, 2022, from https://www.kaggle.com/competitions/tmdb-box-office-prediction