

# Estimering av filminntekter

Daniel Stuberg, 1.11.2023

## BESKRIV PROBLEMET

### SCOPE

Prosjektet skal estimere hvor store inntekter en film vil generere, basert på tilgjengelig informasjon før lansering. I utgangspunktet vil dette beregnes manuelt basert på tidligere historikk, og med maskinlæring kan dette automatiseres og modellen vil kunne ta langt flere hensyn for å estimere inntektene.

Dette vil kunne gi produsenter, distributører og markedsansvarlige et bedre beslutningsgrunnlag for investeringer.

### METRIKKER

Det er viktig at løsningen gjør bedre estimater enn de tradisjonelle metodene.

For å evaluere om løsningen fungerer godt skal Root Mean Squared Log Error (RMSLE) brukes. Det er enorme forskjeller mellom inntektene til de største Hollywood-lanseringene og småfilmer, så det er nødvendig å sikre seg at modellen ikke underprioriterer de mindre filmproduksjonene.

## DATA

Det er flere filmdatabaser hvor store mengder av data er tilgjengelig. Dette vil være data som budsjett, sjanger, skuespillere og utgivelsesdato.

Mange av filmene vil ha unike egenskaper, så det vil være et behov for mye data for å kunne finne reelle mønster.

For å bruke dataene må de struktureres på en måte som er egnet for maskinlæring. Videre vil opprydding i data og feature engineering bli utført som nødvendig. Det er også problematisk at en del filmer vil mangle enkelte data, så det må også tas hensyn til og vurderes om denne type data vil forbedre modellen.

## MODELLERING

Random Forest Regressor skal benyttes i prosjektet. For å finne baseline-ytelse skal det brukes en modell basert på budsjett og popularitet, to features som sannsynligvis har stor innvirkning på inntektene til en film.

Videre må det undersøkes om hver ny feature har et reelt forhold til inntekter og faktisk forbedrer modellen.

## DEPLOYMENT

Den ferdige modellen skal gjøres tilgjengelig for at interessentene skal kunne ta bedre beslutninger knyttet til investeringer og markedsføring.

Etter at modellen er satt i drift vil datagrunnlaget være i konstant endring sammen med filmindustrien. Det er derfor nødvendig å trene modellene på nytt når de ikke lenger har den ønskede treffsikkerheten.