# Box office prediction model

Johan Sørlie, Markus Pedersen og Jørgen Toppen Moen, 17/11/2022

## BESKRIV PROBLEMET

### SCOPE

Prosjektets mål er å etablere en maskinlæringsmodell som kan nøyaktig anslå “box office” inntjening til en film gitt noen parametre (budsjett blant annet). Modellen skal kunne brukes i et interaktivt brukergrensesnitt. Dette vil i første omgang være en forenkling, hvor brukere kan fylle inn noen verdier for og deretter får anslått inntjening.

Dette er ment som et “proof-of-concept”. Dersom den forenklede modellen oppnår suksess, vil det være aktuelt å utvide den til profesjonell bruk. Filmselskaper og kinoer vil kunne dra nytte av modellen i sin budsjettering og planlegging. Kanskje kan modellen og signalisere til filmselskaper hvilke filmer de bør satse på. I dag antar vi at filmselskaper og andre interessenter bruker manuelle prosesser for å anslå inntjening til filmer.

Grundig analyse av arbeid, salg og resultat er viktig for å drive en virksomhet. Ulike faktorer spiller en rolle når ytelsen skal måles via forretnings målinger. Noen av disse kan være salgsinntekter og markedsføring inntekter . En bedrift er avhengig av inntekter for å være en del av et marked. Å spå inntekter er en viktig kjernefunksjon. God markedsføring er også viktig for å oppnå reelle mål. Dette kan gi generell ytelse til sosiale kanaler, reklame og kampanjer.

Foreløpig er det ingen stakeholders i prosjektet annet enn de 3 team-medlemmene som investerer sin tid i å utvikle systemet. Senere er målet å ha flere stakeholders som filmproduksjonsselskaper, produksjonsland hvor filmen blir laget og publikum.I prosjektet finnes det primære og sekundære interessenter. Filmproduksjonsselskaper og produksjonsland er primære som direkte involvert i prosjektet, mens publikum er sekundære.

### METRIKKER

Business metrics er  mest effektiv når det sammenlignes mot benchmark(referansepunkt) eller forretningsmål. Benchmark kan for eksempel være ytelse som spors på tvers av alle områder av bedriften. Minimum business metric for vårt prosjekt er at vi klarer å utvikle produktet for under 300 000 kroner og at vi klarer å selge å produktet for over 2 millioner kroner. Altså en net-profit på minimum 1,7 millioner kroner.

Metrikker som skal brukes til å måle sukssessen til modellen er mean squared error. Det er ennå ikke bestemt hvor liten denne bør være. Dette vil bli avdekket ved å spørre eventuelle kunder hvilke behov de har for nøyaktighet, etter at vi har vist dem vår proof-of-concept.

# DATA

Dataen som skal brukes til å trene modellen er et datasett med 3000 filmer fra The Movie Database (TMDB). Target label for dataen er “revenue”, altså inntjening. Alle attributter er listet opp under. En interessant observasjon er at mange av dem ikke er numeriske. Det vil bety mye opprydding og “engineering” av dataen.

- id
- belongs_to_collection
- budget
- genres
- homepage
- imdb_id
- original_language
- original_title
- overview
- popularity

Vi endte opp med å gjøre om en del av attributtene som ikke var tall, om til numeriske verdier. For produksjonsselskaper lagde vi en attributt som var 1 dersom filmen var produsert av et populært prod.selskap, og 0 ellers. For sjanger one-hot-encodet vi de mest populære sjangerne. For skuespillere (cast) lagde vi en liste over de 40 skuespillerne som deltok i flest filmer, og lagde en ny attributt som anga antall skuespillere i filmen som var i denne listen.

Noen attributter ble fjernet, mye av hensyn til at vårt prosjekt er et proof-of-concept. Brukere skal kunne interagere med modellen for å prøvde den, og da ble det tidlig klart at det var uaktuelt at brukeren måtte fylle inn alt for store mengder data. For eksempel viste våre undersøkelser at det var god korrelasjon mellom antall skuespillere og inntjening, og antall “crew”-medlemmer og inntjening. Dessverre vil ikke dette være nyttig for vår scope, da brukere neppe er villige til å fylle inn hundrevis av navn.

En mulighet for forbedring hadde vært å hente inn mer data ved hjelp av for eksempel imdb_id. Kanskje kunne vi avdekket nye og gode attributter ved å hente ut data fra imdb.

# MODELLERING

Vår plan er å raskest mulig undersøke de ulike attributtene i datasettet, og bestemme om de skal brukes eller ikke. Der tok vi som nevnt noen beslutninger basert på at modellen vår skal kunne brukes i et enkelt brukergrensesnitt. Vi besluttet også å ikke bruke tid i starten på å drive med ulike former for tekstanalyse på de attributtene som inneholdt tekst og nøkkelord.

Modellen vil måtte gjøre regresjon, og basert på tidligere erfaringer virket det aktuelt å teste ut hvordan Random Forests og Gradient Boosting Trees kunne prestere. I valg av viktige attributter og feature engineering tok vi stor inspirasjon fra følgende modell:

https://www.kaggle.com/code/artgor/eda-feature-engineering-and-model-interpretation

## DEPLOYMENT

Planen er å deploye modellen i form av et brukergrensesnitt hvor en begrenset mengde med data kan fylles inn av brukeren. Vi undersøkte løsninger med Flask og Voila, men endte opp med å gå for en løsning som heter Anvil. Anvil er et system for å bygge og deploye web-apps kun ved bruk av Python kode. Gjennom pakker kan man sette opp funksjoner direkte i en notebook som kan fungere som et slags “end-point”. Ved å lage et enkelt brukergrensesnitt i Anvil.works sitt byggemiljø kan man gjøre kall til dette “end-pointet”

Det er i dette “end-pointet” at modellen vi har trent skal brukes til å gjøre predictions på gitt data. Modellen er allerede trent og lagret, og metoden .predict(nydata) brukes på den nye dataen. Den nye dataen (som kom fra brukeren) må selvfølgelig behandles gjennom en pipeline som gjør den klar til å “dyttes” gjennom modellen.

## REFERANSER


Dataset er fra denne konkurransen på kaggle:

https://www.kaggle.com/competitions/tmdb-box-office-prediction

Vår notebook inspirasjon:
https://www.kaggle.com/code/artgor/eda-feature-engineering-and-model-interpretation

Anvil:

https://anvil.works/blog

What are Business Metrics? | Klipfolio