Prognose for inntekt av Kino billetter

[Trym Birkelund Gallefoss, Oskar Windelstad], [15.11.2024]

*Malen inneholder en del beskrivende tekst om hva dere kan skrive om under hver overskrift. Dette er kun forslag til elementer: dropp det som ikke er relevant i ditt tilfelle, og legg til element du finner relevant. Det eneste kravet er at beskrivelsen er i tråd med livssyklusen til maskinlæringsprosjekt skissert i kurset. Det vil si, at prosjektbeskrivelsen følger strukturen reflektert i overskriftene nedenfor. Slett alle instruksjoner (tekst i kursiv) før innlevering.*

**BESKRIV PROBLEMET**

**SCOPE**

Business objective

- Målet med dette prosjektet er å utvikle en maskinlæringsmodell for å forutsi billettinntektene til en film basert på budsjett, popularitet og spilletid. Å skape en nøyaktig prediksjonsmodell kan hjelpe filmselskaper å budsjettere, planlegge markedsføring og forutse suksess

Nåværende løsninger

- I dag baserer mange seg på erfaring og enkle modeller. Maskinlæring kan gi mer presise modeller ved å analysere mønstre fra store mengder data

Business metrics

- Ytelsen måles som nøyaktighet i predikasjon av billettinntekter ved bruk av R^2-score og Mean Absolute Error (MAE)

Pipeline

- Modellen kan integreres i et større system for filmanalyse, som inkluderer datainnsamling, visualisering og rapportering

Stakeholders

- Filmselskaper, investorer, produsenter og markedsførere er de primære interessentene

Tidslinje

- Forberedelse (dag 1-2)
  Prosjektet startet med noen generelle forberedelser og målsetting for hva som ville oppnås. Dette inkluderte sette opp Github og google docs.

- Data samling, «cleaning» og utforskning (dag 3-5)
  Dette var starten på selve prosjektet. Her samlet vi nødvendig data fra film databasen (TMBD). Vi fjernet unødvendig data og så på data som hadde viktige korrelasjoner

- Utvikling av modell (6-7)
  Dette steget så vi på ulike modeller som passet oppgaven.

- Web-app (dag 8-9)

Her satt vi opp en Web-app for å kunne sette inn egne verdier for at modellen kunne gi en predikasjon på hva inntekten ville bli.

Ressurser

- Datasett, PC, python biblioteker (scikit-learn, Gradio)

**METRIKKER**

Business metric ytelse

- Den minimale «Business metric»-ytelsen som kreves for at dette prosjektet skal anses som vellykket, er forbedring av lønnsomheten gjennom økt nøyaktige prognoser. Et spesifikt mål kan være å forbedre presisjonen i inntektsprognosene med minst [X]%, noe som kan forventes å føre til en bedre budsjettallokering og økte inntektsmarginer. Forretningsmålet er å bruke prediktive innsikter til å drive strategiske beslutninger, redusere økonomisk sløsing og utnytte markedsmuligheter.

Maskinlæringsmetrikker

- R^2-score: Måler hvor godt modellen forklarer variasjonen i dataene
- Mean Absolute Error(MAE): Gjennomsnittlig feil i prediksjoner
- Feature importance: for å forstå hvilke variabler som er viktigst for modellen

**DATA**

- **Datakilder:** Datasettet er basert på informasjon om filmer, inkludert budsjett, popularitet og spilletid. Labels er faktiske billettinntekter.
- **Datamengde:** Nåværende datasett inneholder 7,398 filmer hentet fra film databasen (TMDB)
- **Label-kvalitet:** Labels er basert på faktiske rapporterte inntekter, noe som sikrer høy nøyaktighet.
- **Etiske hensyn:** Ingen sensitiv data er i bruk. Datasettet består av kun offentlige tilgjengelig data
- **Representasjon:** Dataene representerer numerisk, og funksjonene skaleres ved behov
- **Feature engineering:** Nye funksjoner som kostnad per minutt eller popularitet per budsjett kan være relevante.
- **Rengjøring:** Manglede verdier fylles inn med median, og unormale verdier fjernes

**MODELLERING**

Modeller

- Første modell: Decision Tree
- Hovedmodell: Random Forest Regressor for bedre ytelse og robusthet

Baseline ytelse

- En enkel lineær modell gir en baseline med en R^2 score på 0.5.

Feature importance

- Viktige variabler identifiseres, som budsjett og popularitet.

Feilundersøkelse

- Store feil analyseres for å finne mønstre som kan forbedre modellen

Iterasjoner

- Modellen finjusteres ved hyperparameter-tuning.

**DEPLOYMENT**

- Bruk: Modellen integreres i et Gradio-grensesnitt for brukervennlige prediksjoner
- Monitorering: overvåk ytelse ved bruk av sanntidsdata og juster modellen ved behov
- Vedlikehold: Planlegg regelmessige oppdateringer med nye data.
- Videre forbredning: Utforsk dypere modeller som Neural Networks eller bruk av ensemble metoder

**REFERANSER**


- ChatGPT
- Hands-on maskinlæring med SciKit-Learn
- Andre referanser inkluderer ulike Youtube-videoer, forum og løsninger fra kaggle konkurranser.