[Spaceship Titanic]

[Anders-Marius Gjerdalen, Adrian Birkedal, Anders Lerang], [03.11.2022]

# **BESKRIV PROBLEMET**

Året er 2912. Vi har fått i oppgave til å løse et kosmisk mysterium. Romskipet Titanic var en interstellar romferje som ble utplassert for en måned siden. Med nesten 13,000 passasjerer om bord, skulle romskipet transporterte emigranter fra vårt solsystem til tre nye habitater.

Romskipet kolliderte da det nærmet seg sin første destinasjon. Romskipet traff en anomali gjemt i en støvsky. Skipet forble intakt, men halvparten av passasjerene ble transportert til en alternativ dimensjon.

For å hjelpe å redde mannskap og passasjerer, skal vi forutse hvilke passasjerer som ble transportert av anomalien ved å bruke data som er gjenvunnet fra skipets datasystemer.

Dette er en Supervised-learning oppgave og et binært klassifiseringsproblem, siden vi skal finne ut om passasjerer ble transportert eller ikke. Ytelsen på prosjektet regnes ut basert på vår «classification accuracy». Dermed får vi anslått en score på hvor treffsikker løsningen vår er opp mot “out of sample data”.

Løsningen kan senere brukes på kommende reiser med romskip med samme personopplysning og data som Space Titanic hadde. Det vil brukes til å forutse hvilke av passasjerene som er størst sannsynlig å bli transportert om romskipet treffer en lik anomali.

Hoved interessenten til prosjektet er White Star Line, som er eiere av romskipet Space Titanic. Andre etater og organisasjoner som driver med romfart vil være indirekte interessenter

**METRIKKER**

Vi antar at prosjektet vil koste rundt 200 000 NOK. Ved å selge maskinlærings systemet som en tjeneste vil “net profit” øke. White Star Line blir kjøperne, men brukere blir passasjerene. De får tilgang til en nettside hvor de kan skrive inn sine personopplysninger, de får her ut et svar på om de vil bli transportert eller ikke med en 80% sannsynlighet.

# **DATA**

Siden problemet er binær klassifisering, blir transported vår label til problemet. Data inneholder informasjon om alle passasjerer, men siden dataen var hentet fra et skadet datasystem er det verdier som mangler. Prosjektet handler i stor grad om å rense denne dataen. Hvor nøyaktig dataen blir renset vil være helt avgjørende for modellens ytelse. Dersom det er høy korrelasjon mellom featurene, kan vi bruke noen features til å fylle inn andre. Det gir også mening at passasjerer som benytter seg av cryosleep ikke bruker særlig penger på kjøpesenteret. Vi vil også måtte komme til å konstruere noen nye features. Eksempelvis er Cabin i formatet (DEKK/KABIN/SIDE). Dette gir mening å dele opp slik at vi har tre nye features. Det samme vil vi også gjøre for passengerId, som forteller noe om gruppennummer og ID innenfor gruppen.

Hvis det finnes mange outliers i datasetet, vil vi muligens skalere dataen. Dette vil gjøre det lettere for modellene og ta hensyn til disse. Det vil sannsynligvis være outliers i bruksvanene til passasjerer.

# **MODELLERING**

Klassifiseringsoppgaver har et godt utvalg av algoritmer å velge i. Ved å shortlisten noen få algoritmer kan vi planlegge hvilken type algoritmer vi vil trene dataen vår på.

Short List:

* LogisticRegression
* SGDClassifier
* GaussianNB
* KneighborsClassifier
* CatBoostClassifier
* LGBMClassifier
* XGBClassifier

Om “predictoren” er dårlige til å lære (“weak learners”) kan det vurderes å bruke en esamble metode.

# **DEPLOYMENT**

Vi kommer til å legge modellen ut på nett ved hjelp av python sitt mikrorammeverk, «flask». Her vil prediksjonene fra modellen brukes til å forutse om en bruker blir transportert basert på inputs. Siden modellen ikke har behov for kontinuerlig ny data, er det naturlig å bruke batch-learning. Det vil kun være nødvendig å trene modellene noen få ganger i året slik at den ikke råtner.

Deployment er ikke det siste steget i maskinlærings syklusen. Etter lansering må systemet må monitoreres. Ved å monitorere kan vi kan unngå situasjoner som gir oss en “negative buisness value”. Eksempler på dette kan være systemsvikt og “model drift”. Vi kommer hovedsakelig til å monitorere ytelse til modellen i forhold til inputs, “prediction” resultater og hva som skjer når modellen lærer ny data i produksjon.

Planen videre er å implementere en funksjon som tar inn hele nye datasett med passasjerer og mannskap. Da kan kunder som har innvisitert i tjenesten ha mulighet til å forutse hvilke av deres passasjerer som transporteres om de treffer en tilsvarende anomali.

# **REFERANSER**

*List opp kilder du har brukt til å utarbeide prosjektbeskrivelsen. Listen av referanser bør sannsynliggjøre at prosjektet kan lykkes (“feasability”)*

[*https://www.kaggle.com/competitions/spaceship-titanic/data?select=test.csv*](https://www.kaggle.com/competitions/spaceship-titanic/data?select=test.csv)

<https://www.kaggle.com/code/samuelcortinhas/spaceship-titanic-a-complete-guide/notebook>

<https://medium.com/@sanidhyaagrawal08/machine-learning-project-checklist-134043501855>