Titanic – Playing the whole game

Andrea, Torstein, Magnus, 18. November

SCOPE

Målet for prosjektet er å konstruere en maskinlærings modell som predikerer hvilke passasjer som overlevde Titanic-forliset. Modellen skal brukes til å lage et videospill som baserer seg på om spilleren hadde overlevd eller ikke. Dette baseres på spillerens sosiale status og eventuelle andre faktorer som kan bli relevant jo mer en utforsker datasettet og de forskjellige prediksjonene som blir gjort.

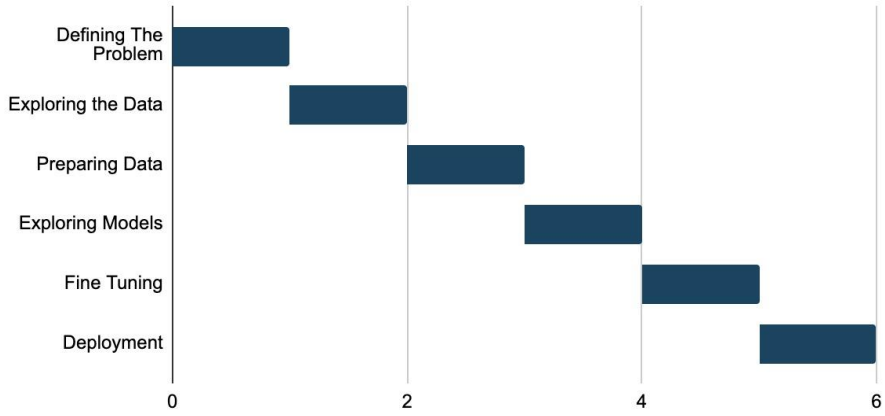
Selv om det er noen tiår siden forliset til Titanic hvor 1502 av de 2224 passasjerene ombord mistet livet grunnet ulike faktorer. Undersøkelser og forskning har i senere tid funnet ut at ikke bare var rene tilfeldigheter rundt hvem som overlevde og hvem som omkom, men hvilken "gruppe" skråstrek hvilken sosial status du hadde, hadde en innvirkning på dine sjanser for å overleve forliset. For å bistå blant annet forskning, skal dette prosjektet i samarbeid med spillselskapet TitanicGames lage en modell som predikere om en spiller overlever eller ei, basert på hvilken gruppe du tilfører.

Så vidt vi har kjennskap til, så eksisterer det ingen tilsvarende løsninger på hvordan man predikere hvem som overlevde forliset hvor deres status blir sett på som en avgjørende faktor. Den eneste, og naturlige måten, man kan se om en person har overlevd, er at man ser på papirene som ble hentet fra etterforskningen og sammenligner dem med dem som var registrert som passasjer på jomfruturen til New York. Denne løsningen baserer seg på tegninger som viser hvor de som overlevde befant seg og hvor eventuelt det var høy sannsynlighet at du overlevde basert på posisjonen til de som overlevde, altså hvor de overlevende befant seg.

Dersom man skulle gjort det uten maskin læring og gjort oppgaven manuelt, ville man ha trengt ett større og variert utvalg data, både fra før og etter forliset. Ved å utføre oppgaven manuelt ville man brukt posisjonen til hvem som overlevde og kunnskap om hvor de befant seg i etterkant på hvor de befant seg da redningsaksjonen var påbegynt, eksempelvis om de var i det iskalde Atlanterhavet eller om de var i en livbåt.

Når det gjelder måling av ytelsen av prosjektet målt med business metrikker, så vil det ikke være enormt relevant for dette prosjektet. Ettersom dette er et prosjekt som skal brukes som en ressurs til å utvikle spillet, som igjen skal brukes til forskning, vil det ikke være noen finansielle tap sånn sett. Grunnet at det ikke ville være noen finansielle tap, er fordi at forskning er ment som en investering som ikke alltid bærer frem store gevinster. Dersom modellen predikter feil, vil ulike metrikkene gi oss et slags forvarsel om at dens treffsikkerhet i prediksjonene er lav eller feile, og vi kan bruke dette til å gjøre eventuelle forbedringer og endringer. Samtidig som at prosjektet er ment for et dataspill og dersom man predikerer feil, så vil ikke konsekvensene være “enorme”. Den “verste” konsekvensen som man kan komme er at man har en urettferdig fordeling om hvem som taper og hvem som vinner i spillet, eksempelvis man bare taper eller bare vinner i spillet.

Prosjektets stakeholders er spillselskapet Titanic Games som skal lage dataspillet. Selskapet trenger informasjon om overlevelses sjansene under forliset og de trenger en modell som kan gi dem informasjon om hvilke personer som overlevde og hvem som ikke gjorde det. Spillet starter ved at man skal lage en avatar og underveis i spillet blir man gitt forskjellige oppgaver. Basert på disse oppgavene og ulike andre parametere så kan man avgjøre om en spiller vinner, altså overlever, eller omkommer, ergo spilleren taper spillet.



Figur 1: Tidslinje for prosjektet, i uker

Prosjektet har noen begrensninger blant annet at disse dataene som er gitt av forskerne, er det ingen garanti for at de er rette opplysninger. Denne mistanken backes opp av at på denne tiden (og som vi skal se på datasettet senere) har en del opplysninger som mangler eller som er feil oppgitt. Dette kan være mange grunner til at det er slik, hvor den mest sentrale er at man ikke hadde særlig god kontroll på personopplysninger på den tiden, ettersom alt var skrevet på papir og det er ikke alltid at all håndskrift er lesbart.

METRIKKER

For å måle kvaliteten på løsningen brukte vi "mean squared error" (MSE) for å måle hvor ofte modellen predikerte rett om passasjeren overlevde eller ikke. Vi vil selvfølgelig ha en modell som gir så bra predikat som mulig og med MSE leter vi etter så lavt tall som mulig. Dette ble brukt når vi evaluerte modellene underveis.

DATA

Dataene som vi har fått er basert på personopplysningene som personalet som jobbet om bord på Titanic hadde for å kunne identifisere hver og enkel passasjer og kunne skille alle passasjerene fra hverandre. Disse opplysningene inneholder blant annet navnet til personen, kjønn, alder, hvilken klasse de hadde bestilt og eventuelt viss de hadde bestilt cabin, i så fall hvilke, og annen info som vi kommer tilbake til senere når vi skal beskrive datasettet.

Kommentert [AS1]: finn ut hva som menes med labels

Datatypene er litt forskjellig, noen er i form av booleans, noen er i form av integers mens andre er bare definert som objekter.

Dataen er hentet fra en Kaggle konkurranse som ligger på Kaggle.com/competitions/titanic . Denne konkurransen er ment som en introduksjonskonkurranse for personer som nettopp har begynt med maskin læring. Dataen er også mulig å få tak i andre varianter, men det krever litt mer grundig arbeid i og med at man må lage datasettet selv. Har valgt å bruke dataen og submitinger til en Kaggle-konkurranse for lettere å kunne se nøyaktigheten på predikasjonene som blir gjort og for å se hvor nøyaktig prediksjonene er i forhold til ny data.

Kommentert [AS2]: legg inn lenke til konkuranse siden på kaggle i referanser men også her som hvorfor ikke

Datasettet vi bruker for å utvikle modellen inneholder personlig informasjon på over 2000 personer, så personvernhensyn er noe vi må tenke gjennom under prosjektet. Modellen er skapt med denne personlige informasjonen, men blir ikke vist videre når vi deployerer modellen. Det andre hensynet er personlig informasjon fra eventuelle sluttbrukere.

Det ble gjort en rekke steg under preprosseseringen, vi slettet flere rader som var unødvendig, som "Name", “Cabin”, og vi brukte One Hot Encoding på “Sex”. Vi brukte heatmaps for å se korrelasjonen mellom de forskjellige featursene. Vi brukte Standard Scaler for å skalere dataen til modellen.

MODELLERING

Vi testet ut flere maskinlæringsmodeller under modelleringsfasen, men endte opp med en ensamble learning modell, sammensatt av ExstraTreeClassifier, og Random Forrest Classifier. Se kode på github:

https://github.com/h584980/ML-Oblig-2---Gruppe-1

For å undersøke “feature importance” brukte vi correlation matrix.

DEPLOYMENT

Vi lagde en nettside med Flask for å sette modellen i drift, her kan man sette inn egne verdier for de forskjellige parameterne som modellen bruker, som kjønn, navn, alder osv... Man trykker da videre og parameterne blr sendt gjennom preprosesseringen og blir så kjørt i maskinlærings modellen. Det blir da gitt et predikat til brukeren på en ny side.