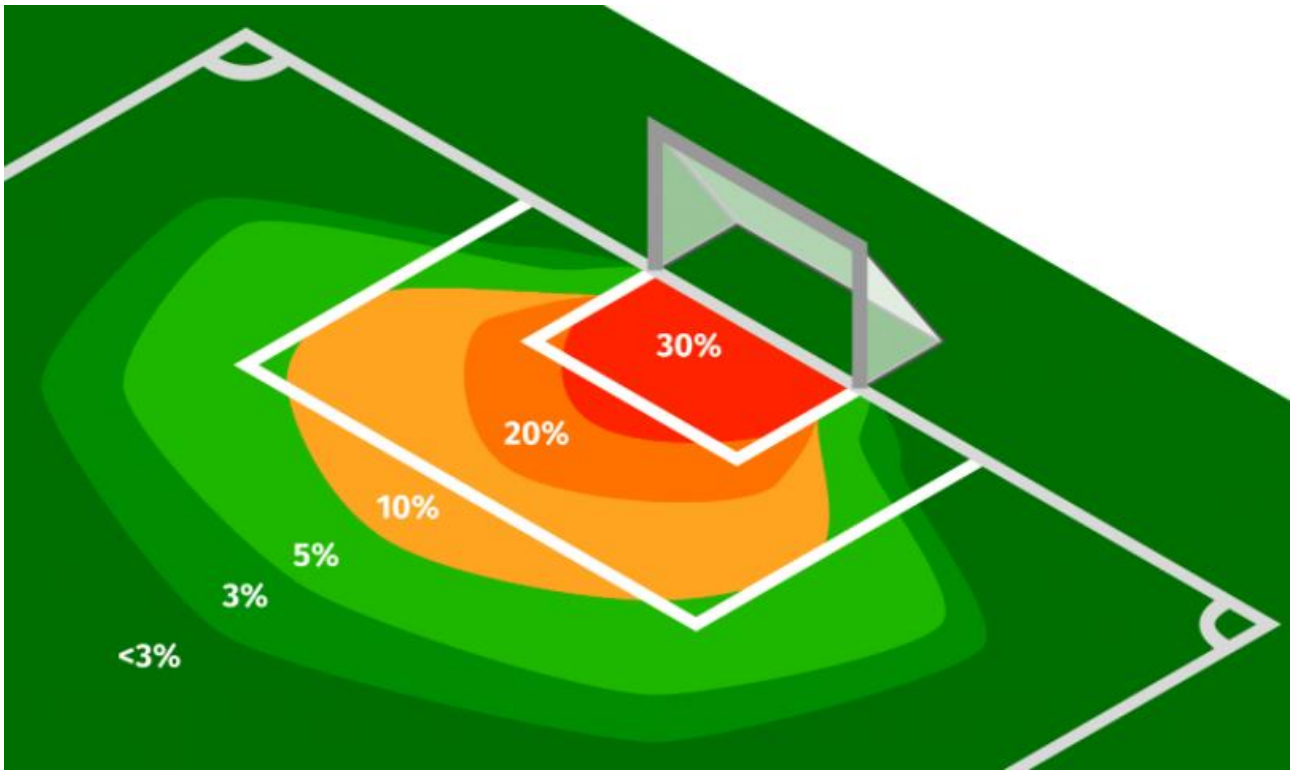
Datadrevne beslutninger i Brøndby IF



Afleveringsfrist:

Mandag den 15. april 2024,

som gruppeaflevering i Wiseflow

Indhold

Generel information om eksamen

Opgaven er en gruppeeksamen, hvor jeres besvarelse skal afleveres som en gruppeaflevering I Wiseflow onsdag den 15. april og mundtlig gruppeeksamen torsdag den 18. april og fredag den 19. april. Den mundtlige udprøvning vil foregå ved en fælles gruppepræsentation efterfulgt af en individuel samtale på 20 minutter. Der vil være individuel evaluering og karakter efter 7-trinsskalaen.

Gruppestørrelsen er 2-4 studerende pr. gruppe, medmindre andet er aftalt med underviser.

Vejledning i eksamensperioden offentliggøres på Moodle.

Til arbejdet med eksamensopgaven ligger der datafiler i samme mappe som eksamensopgaven.

Evt. spørgsmål stilles til adjunkt Thorbjørn Baum eller adjunkt Thorbjørn Wulf.

Opgave 1 – Expected Goals Model (xG model)

En af hjørnestene inden for fodboldanalyse er xG – Expected Goals – der estimerer sandsynlighed for at et givent skud bliver til ét mål.

I skal benytte data for Superligaen 2022/2023, der ligger I eksamensmappen.

Opgave 1.1 – Opdeling i trænings- og testdata for skud

Indlæs filerne for skud i kampene i Superligaen for sæsonen 2022/2023 og opdel jeres data i træning og test. Forklar jeres overvejelser i forhold til opdelingen i træning og test givet data for skud i kampene.

Opgave 1.2 – Forklarende variable, forklaring og grafiske illustrationer beskrivende statistik

Giv en kort beskrivelse af de forklarende variable som I ønsker at benytte i en xG model. (Hint: beskriv grundigt jeres feature engineering). Derudover skal I lave grafiske illustrationer af de ønskede forklarende variable, samt beskrivende statistik.

Opgave 1.3 – Forklarende variable og effekt på om et skud bliver til mål

Hvilke af jeres valgte variable vil I forvente har en effekt på om et skud bliver til et mål i Superligaen 2022/2023? (Giv en forklaring på, hvorfor I mener netop disse variable vil have en betydning).

Opgave 1.4 – Forudsige om et givent skud bliver til et mål

Opstil en klassificeringsmodel til forudsigelse af om et givent skud bliver til et mål. I skal benytte trænings- og testdata til at udvælge jeres model. Giv en forklaring på den valgte model (hint: det forventes I afprøver forskellige modeller).

Opgave 1.5 – Konklusioner på xG model

Lav en præsentation af jeres model fra opgave 1.4. I præsentationen skal indgå minimum to grafiske illustrationer og to tabeller (Vær opmærksom på I bliver bedømt inden for faget visualisering). I må gerne starte jeres præsentation med et resumé af jeres resultater.

Opgave 1.6 – I virkeligheden

Hvad kan Brøndby IF bruges jeres model og resultater til? (Hint: Test jeres model på data for den indeværende sæson.)

Opgave 2 – Expected points model (xP model)

Denne opgave er stillet i samarbejde med Nicolai Fernandez Pedersen, Ph.D., Head of Football Data Analytics i Brøndby IF. Derfor skal opgave løses for Nicolai til brug for Brøndby IF. Det forventes den studerende selvstændig finder information om modellen Expected Points (xP), men benytter data stillet til rådighed for eksamen

Opgave 2.1 – Opsætning af model for Expected Points (xP)

Baseret på xG i data for Superligaen 2023/2024 skal I udregne Expected Points (xP) for os. Dette kræver, at I simulerer kampene et passende antal gange for at finde sandsynligheden for forskellige udfald i de enkelte kampe. Når det er gjort, kan I udregne, hvor mange point vi burde have.

Opgave 2.2 – Validering af data for xP model

Efter indførelsen af Video Assistant Referee (VAR) fløjter dommerne ikke længere tvivlsomme offsides, men venter på, at spilsekvensen er overstået, før de eventuelt fløjter for offside. Dette gøres for at sikre, at hold ikke utilsigtet fratages mål, som er tæt på offsidegrænsen. En utilsigtet konsekvens er, at vi nu registrerer flere skud i statistikkerne, da offside ikke nødvendigvis bliver fløjtet, selvom der på den ene eller anden måde var offside i opspillet. I bedes derfor validere data for skud i følgende tre kampe:

1) Brøndby IF – Silkeborg, den 17. marts 2024
2) Viborg – Brøndby, den 10. marts 2024
3) Brøndby IF – Vejle BK, den 3. marts 2024

Giv jeres vurdering af om den utilsigtede konsekvens af VAR, som Brøndby IF er nervøs, vil have en signifikant effekt på jeres xP model. Vær opmærksom på I alene har valideret 3 kampe.

Opgave 3 – Overblik over spillere i Holland og Polen

I denne opgave skal I hjælpe Brøndby IF med at få et overblik over spillere i den hollandske og polske liga. I **kan**, hvis ikke andet er skrevet i opgaveteksten, benytte eventdata fra Wyscout for forrige sæson 2022/2023, samt sæsonen 2021/2022. I gør selv de nødvendige antagelser i forhold til sæsonerne.

Brøndby IF er mest fokuserede på afleveringer og skud, når de skal danne sig et overblik over spillere i de to ligaer. I må dog gerne benytte andre typer af events, hvis I mener det relevant for overblikket.

Jeres opgave er alene informativ og Brøndby IF ønsker alene et overblik over spillere i de to ligaer for at øge deres kendskabsgrad til spillere i Holland og Polen. Opgaven skal derfor ikke ses som Brøndby IF, der ønsker at erstatte nuværende spillere i truppen.

Opgave 3.1 – Afleveringer

Lav en Clustering model af afleveringer for henholdsvis den hollandske og den polske liga. I bestemmer selv antallet af clustre. Hvad kendetegner jeres clustre som modellen valgt at opdele afleveringerne på? (Hint: sæt en sigende overskrift på hvert af jeres clustre – husk det skal give Brøndby IF et overblik)

Opgave 3.2 – Afslutninger (skud)

Lav en Clustering model af afslutninger i ét datasæt for henholdsvis den hollandske og den polske liga. I bestemmer selv antallet af clustre. Hvad kendetegner jeres clustre som modellen valgte at opdele afslutninger på? (Hint: sæt en sigende overskrift på hvert af jeres clustre – husk det skal give Brøndby IF et overblik)

Opgave 4 – Præsentation til Brøndby IF

Udgangspunkt for denne opgave er jeres resultater i opgave 3. Det vil sige, at det ikke er nødvendigt at beskrive modellen, samt jeres test og validering af modellen fra opgave 3. I skal alene fokusere på kommunikation af jeres resultater til Brøndby IF.

I skal være klar over jeres rolle fra opgave 3. Hvem er modtager i Brøndby IF af jeres resultater? I skal benytte Shiny til at kommunikere jeres resultater.

Opgave 4.1 – Visualisering af clustre for afleveringer

Lav et Dashboard, der giver Brøndby IF et overblik over afleveringerne i henholdsvis den hollandske og polske liga.

Opgave 4.2 – Visualisering af clustre for afslutninger

Lav et Dashboard, der giver Brøndby IF et overblik over afslutninger (skud) i henholdsvis den hollandske og polske liga.

Opgave 4.3 – Visualisering af clustre for kampene

Lav et Dashboard, der giver Brøndby IF et overblik over afslutninger (skud) i henholdsvis den hollandske og polske liga.

Opgave 5 – ”Beskrivende” statistik og visualisering

I denne del lægges der op til, at I kigger på data med et bestemt blik – nemlig forskelle på mænd og kvinders måde at spille fodbold på. I skal forestille jer, at I arbejder for en organisation, som ønsker at fremme kvindefodbold i Danmark. Det sker som et led i en større plan for at øge kvinders adgang til dele af samfundet, som hidtil har været domineret af mænd. Så organisationen tænker, at hvis man f.eks. kan få flere kvinder til at spille fodbold, vil det gøre dem mere ”robuste” til at tage kampen om på andre områder.

Opgave 5.1 – Kampe i tal

Det er en bunden opgave at finde statistisk belæg for, at der er forskel på mænd og kvinder når det drejer sig om fodbold. I vælger selv variabler fra tilgængelige Statsbombdata men I skal altså kunne nå frem til resultater som visualiserer forskelle mellem mænd og kvinder.

Opgave 5.2 – Freeze-Frame i tal

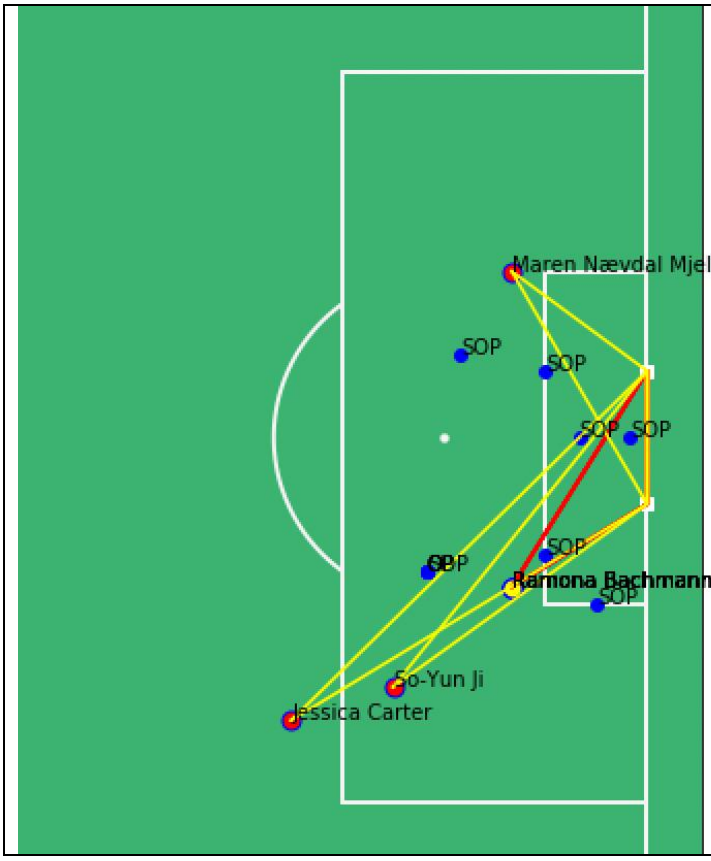
I skal forsøge at underbygge følgende hypotese med dataanalyse:

Vi forventer at finde en større procentvis andel af skud som skulle have været afleveret til en bedre placeret spiller hos mænd end hos kvinder.

Den underliggende antagelse er, at kvinder har et stærkere blik for det sociale end mænd og dermed vil kvinder være bedre til at aflevere i situationer, hvor en medspiller står bedre placeret. I kan selv afgøre, hvordan I vil modellere ”bedre” placering, men minimumskravet er, at I kigger på antal modstandere i den trekant, der dannes mellem personen, der skyder, og de to målstolper og sammenligner med medspillernes trekant.

Analysen skal munde ud i en simpel visualisering af jeres resultat. I skal også formidle den matematik I har benyttet jer af i analysen. Det skal gøres på en måde, så ikke-faglige deltagere i projektet kan forstå det. I skal desuden illustrere jeres undersøgelse med minimum to cases.

Hvis det viser sig, at hypotesen ikke holder, skal I komme med forslag til, hvordan organisationen alligevel kan få hypotesen til at passe ved hjælp af “kreativ” statistik.

	I dette eksempel (eb59c3f1-e28b-4f57-9fe8-0aae82c5ecf6) står Maren faktisk bedre placeret, da hun kun har én spiller i sin trekant, hvorimod hende der tager skuddet (Ramona) har to. I kan vælge at forbedre metrikken ved at inkludere jeres xG på hver medspiller og/eller kigge på, hvor langt der er fra Ramona til Maren og/eller, hvor mange modspillere, der står "tæt på" linjen mellem Maren og Ramona og/eller, hvor mange modspillere, der står i cirklen omkring Maren og Ramona.

Opgave 5.3 – Etikken

I skal nu stille spørgsmålet: Hvor langt vil jeg gå i mine "beskrivelser" i forhold til den agenda som din arbejdsgiver har med analysen. I skal forsøge at se det med konsekvensetiske briller – altså overføre Trolley-diskussionen til dataanalyse ved at opstille en almen regel a la "Hvis min handling redder 5, så er det etisk forsvarlig at slå én ihjel". I skal dernæst se det med Kants øjne ved at overveje om jeres "beskrivelser" vil kunne klare et pligtetisk tjek. Hvilke af de to etiske grundpositioner vil kunne bruges til at forsvare at I f.eks. gik med til at manipulere med data?

Opgave 6 – Videnskabsteori - tæt på virkeligheden

Opgave 6.1 – Hermeneutik – mennesket bag facaden

I skal lave en top-10 blandt kvinderne. I må selv bestemme hvad I måler på. I skal så bruge den information til at tegne et portræt af den spiller I har valgt. Men hermeneutisk - det vil sige en redegørelse for hvordan verden ser ud gennem hendes øjne – hendes livsverden. Det er naturligvis ikke let, men tænk på hvad man f.eks. kan finde på nettet om Nadia Nadim.

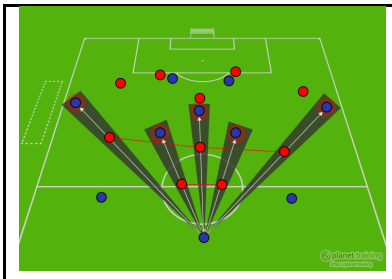
Opgave 6.2 – Logisk positivisme i fodbold

Lav en prioriteret liste over udvalgte features fra *Statsbombs* data. I skal prioritere efter hvor "tætte" de er på "virkeligheden". I kan f.eks. prioritere ud fra en overvejelse over hvordan man ville opdage en fejl, hvordan ville en fejl vil se ud og hvordan ville man i givet fald rette den.

Opgave 6.3 – Hvad ser man når man går helt tæt på virkeligheden?

Med udgangspunkt i trackingdata fra en kamp mellem Vejle BK og Odense BK skal I lave en visualisering af *covering shadows*. I kan enten vælge at tage udgangspunkt i de rå trackingdata (vbob-data.json) eller i den filtrerede fil, hvor data ligger som csv (vbob.csv). Metadata om kamp og spillere ligger i vbob-meta.json. Vær opmærksom på, at de første mange linjer i csv-filen er starten, hvor de giver bolden op og derfor er måske ikke det bedste sted at starte.

Kravene til visualiseringen kommer fra Brøndby IF:



Mikkel Keldmann (former *head of quantitative analysis*, BIF):
For en enkelt pasning kig på pasningsmuligheder ved at implementere en visualisering af "covering shadows" – se billede nedenfor.

Fremhæv "bedste" aflevering baseret på progression mod mål – samt den faktiske aflevering.

Løs følgende delopgaver for at nå frem til et udspil til Brøndby IF på baggrund af Mikkel Keldmanns overvejelser:

1) beskriv data og redegør for forskellen på csv og json, fordele og ulemper

2) redegør for fordele og ulemper ved trackingdata versus eventdata

3) find de frames som dækker det første mål i kampen mellem Odense BK (OB) og Vejle BK (VB).

4) vælg en eller flere relevante frames og plot banen med hold og bold samt navne på spillerne

5) redegør for hvordan trackingdata illustrerer de problemer man får i den logiske positivisme når man skal vise hvordan man kobler komplekse observationer til "atomare" sansedata. Inddrag socialkonstruktivismen som alternativ position - både ontologisk og epistemologisk.