**Expected Goals Superligaen 2021-22**

Denne Expected Goals model er lavet på Wyscout data for Superligaen sæson 21/22 (april 2022).

Datasættet har 2666 observationer af 33 variable, som opdeles tilfældigt i 75% træningsdata og 25% testdata.

**Variable:**

Ud af de 33 variable blev følgende varibale udvalgt til modellen:

Kropsdel: Vi så i vores beskrivende statistik for superligakampe at der er størst sandsynlighed for at skud lavet med hovedet eller andre kropsdele bliver til mål, selvom kun få skud bliver lavet med disse. Vi tror derfor at kropsdel kunne have en mindre betydning for udfaldet af skud.

Halvleg: Vi har før set at flest skud bliver lavet i anden halvleg, men at sandsynligheden for at score mål faktisk er større i første halvleg.

Lokation X, Lokation Y, Distance til mål: : Vi antager at jo tættere e man er på målet, jo større er sandsynligheden for at man scorer.

Hold formation & Modstander hold formation: Gennem vores beskrivende statistik har vi set at det oftest er de samme holdformationer som anvendes. Vi er derfor interesserede i at se om det overhovedet har en betydning for udfaldet af skud.

Vinkel til mål: Vi antager at jo mindre vinklen til målet er, jo større er sandsynligheden for at score.

**Opbygning af model:**

**Classification tree:**

Den første klassificeringsmodel vi forsøgte at anvende i vores xG model er et classification tree. Vi valgte denne model da den både er nem at forklare, men også umiddelbart meget intuitiv at forstå, da den ligner den måde mennesker tager valg på.

Vores classification tree endte med at se således ud.

Diagram

Description automatically generated

Som vi kan se, anvender vores classification tree kun variablerne Lokation Y og Lokation X til at forudsige med.

For at lave et mere overskueligt træ, tjekker vi hvor der kan prunes.

Chart, box and whisker chart

Description automatically generated

Vi kan se at antallet af fejl vil være det samme lige meget hvor lille træet er. Vi skærer det derfor til ved 7 noder, for at få et mere overskueligt træ:

Diagram

Description automatically generated

Her ser vi at der kun er to udfald som giver mål (1).   
Lokation Y er mellem 29,5 og 30,5 samtidig med at Lokation X er mellem 95,5 og 85,95.

Lokation Y er mindre end 49,5 og Lokation X er større end 95,5.

**Random forest:**

Vi har også lavet en xG model med en random forest klassifikationsmodel.

Chart, scatter chart

Description automatically generated

Her kan vi se hvor vigtige modellen vurderer de forskellige variabler. Lokation Y er stadigvæk signifikant, men holdformation er også vigtigt. Vi tester derfor denne model, for at se om den skulle være mere præcis.

## 

## **Konklusioner på xG model**

**Resultat af classification tree:**

|  |  |  |
| --- | --- | --- |
|  | Forudsigelse: Ikke mål | Forudsigelse: Mål |
| Egentlige: Ikke mål | 576 | 72 |
| Egentlige: Mål | 13 | 4 |

Vores classification tree har forudsagt 87,2% rigtigt. Den er dog bedst til at forudsige hvornår der ikke er mål.

**Resultat af random forest:**

|  |  |  |
| --- | --- | --- |
|  | Forudsigelse: Ikke mål | Forudsigelse: Mål |
| Egentlige: Ikke mål | 581 | 65 |
| Egentlige: Mål | 8 | 11 |

Vores random forest model har forudsagt 90,4% rigtigt, og er altså mere præcis end classification tree.

Selvom random forest er bedre til at forudsige, kan et classification tree være nemmere at aflæse og forstå. Man må derfor gøre op med sige selv om de 3% gør en signifikant forskel.