Stockholms universitet

Användning av logistisk regression för att prediktera utfallet i snooker

Leo G. Levenius

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2023:10
Matematisk statistik
Juni 2023

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2023:10**
http://www.math.su.se/matstat

Användning av logistisk regression för att prediktera utfallet i snooker

Leo G. Levenius*

Juni 2023

Sammanfattning

Syftet med det här arbetet är att undersöka hur väl logistisk regression kan användas för att prediktera vinnaren i en snookermatch. Detta görs med hjälp av statistik över spelarna samt resultat från tidigare matcher och turneringar. En mängd möjliga förklarande variabler presenteras som exempelvis ranking, antal vunna matcher, hemland, typ av turnering, prissumma och omgång (final, semifinal, et cetera). Även tvåvägs-interaktioner mellan variablerna undersöks. Modeller tas fram utifrån hur de presterar i AIC, BIC, residualavvikelse samt Hosmer-Lemeshow-testet. Därefter mäts deras prediktiva förmåga hos ett helt nytt datamaterial med hjälp av noggrannhet, sammanblandningsmatriser och AUC.

Resultatet ger flera olika modeller, men den som i slutändan väljs är en modell med bara en förklarande variabel – skillnaden i spelarnas ranking. Modellen hade rätt i sina prediktioner i 60 procent av fallen. Snooker visar sig vara en relativt svårpredikterad sport, jämfört med exempelvis fotboll och hockey, med flera oväntade utfall där den överlägset bättre rankade spelaren förlorade. Modellen är visserligen bättre än vad att godtyckligt gissa vilken spelare som kommer vinna hade presterat, vilket får ses som ett lägsta mått på användbarhet.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: leo.levenius@gmail.com. Handledare: Taras Bodnar & Kristoffer Lindensjö.

Abstract

This study aims to investigate how well logistic regression can be used to predict the winner in a snooker game. This is done using statistics on the players and results from previous matches and tournaments. A range of possible explanatory variables are presented, such as ranking, number of wins, country, type of tournament, prize money, and round (final, semifinal, et cetera). Two-way interactions between the variables are also examined. Models are developed based on their performance in AIC, BIC, residual deviation, and the Hosmer-Lemeshow test. Then, their predictive ability is measured on an entirely new data set using accuracy, confusion matrices, and AUC.

The result produces several different models, but the one ultimately chosen is a model with only one explanatory variable – the difference in the players' rankings. The model was correct in its predictions in 60 per cent of cases. Snooker turns out to be a relatively difficult sport to predict, compared to, for example, football and hockey, with several unexpected outcomes where the significantly better-ranked player lost. The model is at least better than randomly guessing which player would win, which should be seen as the lowest measure of usefulness.

Förord

Jag vill börja med att tacka mina handledare, professor Taras Bodnar och docent Kristoffer Lindensjö, som bidragit med hjälp, idéer och återkoppling under arbetets gång. Jag vill även rikta ett tack stort till fil. stud. Viktor Ellerud för hjälp med att bolla idéer och korrekturläsning.

Denna rapport utgör ett examensarbete om 15 högskolepoäng för en filosofie kandidatexamen i matematisk statistik. Arbetet är skrivet vid Matematiska institutionen på Stockholms universitet.

Innehållsförteckning

1 Introduktion **1**
1.1 Frågeställning . . . 1
1.2 Tidigare forskning . . . 1

2 Teori **2**
2.1 Snooker . . . 2
2.2 Logistisk regression . . . 2
2.3 Newton-Raphson och ML-skattningar . . . 3
2.4 Akaikes informationskriterium . . . 4
2.5 Bayesianska informationskriteriet . . . 4
2.6 Framåtselektion och bakåteliminering . . . 5
2.7 Residualavvikelser . . . 5
2.8 Hosmer-Lemeshow-testet . . . 5
2.9 Sammanblandningsmatris och noggrannhet . . . 6
2.10 AUC och ROC-kurva . . . 6

3 Data **7**
3.1 Urval . . . 7
3.2 Antaganden . . . 9
3.3 Multikollinearitet . . . 10
3.3.1 Variansinflationsfaktor . . . 10
3.3.2 Parvis korrelation . . . 11

4 Metod **12**

5 Resultat **12**

6 Slutsats **14**

7 Diskussion **14**

Referenser **17**

Bilagor **18**

1 Introduktion

Betting är en bransch som omsätter miljarder kronor varje år. En av sporterna som det går att satsa pengar på är snooker. Det är därför av intresse för både företagen som erbjuder bettingen och kunderna som satsar pengar att försöka prediktera hur en match kommer att sluta. Detta kan göras med hjälp av data från tidigare matcher och turneringar som vinstandel, ranking, hemland, et cetera. En metod för att prediktera sannolikheten för ett visst utfall är logistisk regression vilket är det som kommer att användas i den här rapporten.

1.1 Frågeställning

Frågeställningen till detta arbete är ”hur användbart är logistisk regression och den tillgängliga datan för att prediktera utfallet i snooker?”

1.2 Tidigare forskning

Det finns en rik flora av tidigare forskning där forskare har använt logistisk prediktion för att prediktera utfallet i olika sporter. Bland annat finns Willoughby (2002). Författaren använder logistisk regression för att prediktera utfallet i kanadensisk fotboll för tre olika lag med separata modeller för varje lag. De förklarande variablerna – ”rushing yardage”, ”passing yardage”, antal passningsbrytningar, antal fumlingar samt antal ”quarterback sacks” – är med i alla modellerna, men koefficienterna varierar. Variablerna saknar interaktioner. Istället är alla variabler differensen av de respektive värdena för de två mötande lagen. Slutresultatet är tre simpla modeller med fem oberoende förklarande variabler som resulterade i en noggrannhet (antalet korrekta prediktioner dividerat med totala antalet matcher i testdatan) på 79, 86 respektive 90 procent.

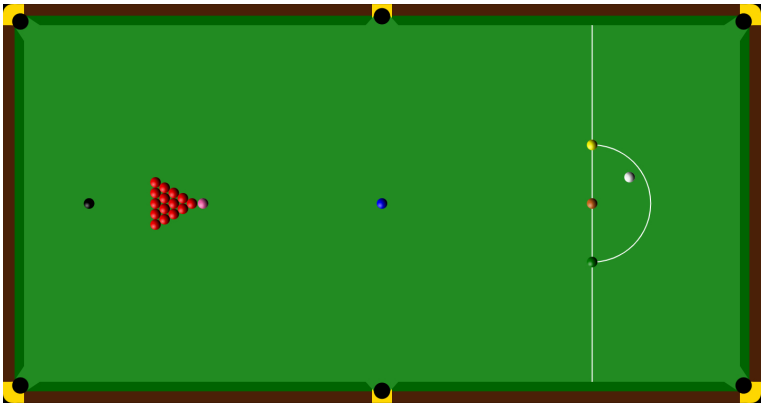
En annan artikel i ämnet är Koo, et al. (2016). Där predikterar författarna utfallet i ishockey med hjälp av logistisk regression och data från tidigare matcher. Bland de förklarande variablerna som används finns antal gjorda mål, hemma- eller bortamatch och antal fouls. Totalt hade modellen rätt i 82 procent av fallen.

När det gäller tidigare forskning gällande prediktering av specifikt snooker är urvalet mycket mindre. En rapport som avhandlar detta är Li, et al. (2021). Författarnas huvudfokus ligger på metoden artificiella neurala nätverk (ANN), men även logistisk regression används (dock med väldigt lite detaljer kring modellarbetet). De variabler som finns med är bland andra prissumma, plats, ranking och nationalitet. Det saknas information om hur variablerna interagerar med varandra samt värden på koefficienterna. Träningsdatan bestod av matcher från de 20 turneringarna varje säsong med högst prissumma över åren 2011–2021. Noggrannheten för logistisk regression var 66 procent, vilket kan jämföras med 71 procent för ANN. Prediktionen utfördes på 33 nya matcher från 2020.

2 Teori

2.1 Snooker

Snooker är en biljardsport uppfunnen av brittiska översten sir Neville Chamberlain under sin tjänstgöring i Indien år 1875. En match spelas mellan två spelare i bäst av n så kallade *frames* (jämför set i tennis). På bordet läggs en vit köboll, 15 röda bollar värda 1 poäng och sex färgade bollar värda 2–7 poäng (se Figur 1). Den ena spelaren börjar med att försöka sänka en röd boll. Lyckas spelaren får han i nästa stöt skjuta ner en färgad boll som sedan återläggs på bordet. Detta upprepas tills spelaren inte lyckas sänka en boll, varav det blir den andra spelarens tur. På bordet finns som mest bollar värda 147 poäng (15 röda, 15 svarta samt en av varje färgad boll). Framet är över när antingen alla bollar sänkts, den ena spelaren leder med mer än 7 poäng när bara den svarta bollen är kvar eller när den ena spelaren gett upp (World Professional Billiards and Snooker Association 2022).



Figur 1: Snookerbord med alla bollar upplagda i respektive startposition (Wikimedia Commons 2021).

2.2 Logistisk regression

Logistisk regression är en statistisk metod som används för att bestämma kopplingen mellan en binär responsvariabel, Y, och olika förklarande variabler, $\mathbf{X} = (X_1, X_2, \ldots, X_{p-1})^t$. Låt sannolikheten för vinst vara $\pi(\mathbf{x}) = \mathbb{P}(Y = 1 \mid \mathbf{X} = (x_1, x_2, \ldots, x_{p-1})^t)$. Vi definierar oddsen för vinst som $\pi(\mathbf{x})/(1 - \pi(\mathbf{x}))$. Enligt Agresti (2013, ss. 119–120) kan logaritmen av oddsen för vinst, logit, uttryckas med det lineära sambandet

$$\operatorname{logit}[\pi(\mathbf{x})] = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \alpha + \boldsymbol{\beta}^t\mathbf{x}, \tag{1}$$

där $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{p-1})^t$. Genom att transformera (1) samt skatta α och $\boldsymbol{\beta}$ (för hur det går till se avsnitt 2.3) får vi en skattning av sannolikheten för det sökta utfallet givet

tidigare data att vara

$$\widehat{\pi}(\mathbf{x}) = \frac{1}{1 + \exp\left\{-\left(\widehat{\alpha} + \widehat{\boldsymbol{\beta}}^t \mathbf{x}\right)\right\}}. \tag{2}$$

Fördelen med denna metod är att den lineära modellen (1) kan anta alla reella värden, medan sannolikheten (2) är begränsad inom intervallet $[0, 1]$. För något $\widehat{\beta}_i > 0$ gäller det att när x_i ökar så ökar $\widehat{\pi}(\mathbf{x})$, givet allt annat lika. Om $\widehat{\beta}_i < 0$ så minskar istället $\widehat{\pi}(\mathbf{x})$ då x_i blir större. Med hjälp av $\widehat{\pi}(\mathbf{x})$ kan vi slutligen prediktera utfallet enligt

$$\widehat{Y} = \begin{cases} 1 \text{ om } \widehat{\pi}(\mathbf{x}) \geq b, \\ 0 \text{ eljest;} \end{cases}$$

där $b \in [0, 1]$. Således gäller det att när brytpunkten b är nära 0 så ökar andelen matcher predikterade att sluta i vinst, $\widehat{Y} = 1$, och tvärsom. Vanligtvis sätts $b = 0.5$.

2.3 Newton-Raphson och ML-skattningar

För att skatta α och $\boldsymbol{\beta}$ använder vi en applicerad Newton-Raphson-algoritm. Den kräver att vi vet score-funktionen samt Fisher-informationen till den bakomliggande funktionen. Vi beräknar dem nedan.

Låt $\theta = (\alpha, \beta_1, \ldots, \beta_{p-1})^t$. Det gäller för en enskild match, k, att $Y_k \sim \mathrm{Be}(\pi(\mathbf{x}_k))$. Således är likelihoodfunktionen för alla n observationer lika med

$$\begin{aligned} L(\theta) &= \prod_{k=1}^{n} \pi(\mathbf{x}_k)^{y_k} (1 - \pi(\mathbf{x}_k))^{1-y_k} = \prod_{k=1}^{n} (1 - \pi(\mathbf{x}_k)) \left(\frac{\pi(\mathbf{x}_k)}{1 - \pi(\mathbf{x}_k)} \right)^{y_k} \\ &= \prod_{k=1}^{n} (1 - \pi(\mathbf{x}_k)) \cdot \exp\left\{ y_k \cdot \ln\left(\frac{\pi(\mathbf{x}_k)}{1 - \pi(\mathbf{x}_k)} \right) \right\}, \end{aligned}$$

vilket med hjälp av (1) kan uttryckas i θ som

$$L(\theta) = \prod_{k=1}^{n} \left(1 + \exp\left\{\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right\}\right)^{-1} \cdot \exp\left\{y_k \cdot \left(\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right)\right\}.$$

Log-likelihooden blir då

$$l(\theta) = \ln(L(\theta)) = \sum_{k=1}^{n} \left(-\ln\left(1 + \exp\left\{\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right\}\right) + y_k \cdot \left(\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right)\right).$$

Vi kan nu beräkna score-funktionen

$$S(\theta) = \begin{pmatrix} \frac{\partial l(\theta)}{\partial \alpha} \\ \frac{\partial l(\theta)}{\partial \beta_1} \\ \vdots \\ \frac{\partial l(\theta)}{\partial \beta_{p-1}} \end{pmatrix} = \begin{pmatrix} \sum_k \left(-\frac{\exp\left\{\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right\}}{1 + \exp\left\{\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right\}} + y_k \right) \\ \sum_k \left(-\frac{x_{1,k} \cdot \exp\left\{\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right\}}{1 + \exp\left\{\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right\}} + y_k x_{1,k} \right) \\ \vdots \\ \sum_k \left(-\frac{x_{p-1,k} \cdot \exp\left\{\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right\}}{1 + \exp\left\{\alpha + \boldsymbol{\beta}^t \mathbf{x}_k\right\}} + y_k x_{p-1,k} \right) \end{pmatrix},$$

samt Fisher-informationen

$$
I(\theta) = \begin{pmatrix} -\frac{\partial^2 l(\theta)}{\partial \alpha^2} & -\frac{\partial^2 l(\theta)}{\partial \alpha \partial \beta_1} & \cdots & -\frac{\partial^2 l(\theta)}{\partial \alpha \partial \beta_{p-1}} \\ -\frac{\partial^2 l(\theta)}{\partial \beta_1 \partial \alpha} & -\frac{\partial^2 l(\theta)}{\partial \beta_1^2} & \cdots & -\frac{\partial^2 l(\theta)}{\partial \beta_1 \partial \beta_{p-1}} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial^2 l(\theta)}{\partial \beta_{p-1} \partial \alpha} & -\frac{\partial^2 l(\theta)}{\partial \beta_{p-1} \partial \beta_1} & \cdots & -\frac{\partial^2 l(\theta)}{\partial \beta_{p-1}^2} \end{pmatrix}
$$

$$
= \begin{pmatrix} \sum_k \frac{\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}}{\left(1+\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}\right)^2} & \sum_k \frac{x_{1,k} \cdot \exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}}{\left(1+\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}\right)^2} & \cdots & \sum_k \frac{x_{p-1,k} \cdot \exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}}{\left(1+\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}\right)^2} \\ \sum_k \frac{x_{1,k} \cdot \exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}}{\left(1+\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}\right)^2} & \sum_k \frac{x_{1,k}^2 \cdot \exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}}{\left(1+\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}\right)^2} & \cdots & \sum_k \frac{x_{1,k} x_{p-1,k} \cdot \exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}}{\left(1+\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}\right)^2} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_k \frac{x_{p-1,k} \cdot \exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}}{\left(1+\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}\right)^2} & \sum_k \frac{x_{p-1,k} x_{1,k} \cdot \exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}}{\left(1+\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}\right)^2} & \cdots & \sum_k \frac{x_{p-1,k}^2 \cdot \exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}}{\left(1+\exp\{\alpha+\boldsymbol{\beta}^t \mathbf{x}_k\}\right)^2} \end{pmatrix}.
$$

Vi har nu allt som krävs för att skatta θ. Vi utför över iterationerna $t = 1, 2, 3, \ldots$ den applicerade Newton-Raphson-algoritmen

$$
\theta^{(t+1)} = \theta^{(t)} + I\left(\theta^{(t)}\right)^{-1} \cdot S\left(\theta^{(t)}\right).
$$

Startvärdet $\theta^{(1)}$ väljs så att $\det(I(\theta)) \neq 0$. När $\theta^{(t+1)} = \theta^{(t)}$ är vi klara och resultatet är en maximum likelihood-skattning, $\widehat{\theta}_{\mathrm{ML}}$ (Held & Sabanés Bové 2020, ss. 31, 163).

2.4 Akaikes informationskriterium

Akaikes informationskriterium (AIC) är ett kriterium för att jämföra olika modellers förmåga att passa ihop med träningsdatan (*goodness of fit*). Kriteriet lyder att för bästa prediktioner välja den modell som minimerar AIC. Det definieras av Sundberg (2022, s. 284) som

$$
\mathrm{AIC} = 2p - 2l(\widehat{\theta}_{\mathrm{ML}}),
$$

där p är antalet parametrar i modellen. AIC straffar modeller med många parametrar till fördel för enklare modeller.

2.5 Bayesianska informationskriteriet

Bayesianska informationskriteriet (BIC) är ett kriterium som fungerar i likhet med AIC. Dess definition är enligt Agresti (2013, s. 212)

$$
\mathrm{BIC} = p \ln n - 2l(\widehat{\theta}_{\mathrm{ML}}),
$$

där n är antalet observationer i träningsdatan. För $n > 7$, vilket är fallet i detta arbete, gäller det för alla modeller att BIC > AIC. Det innebär även att BIC straffar modeller med många parametrar ännu mer än vad AIC gör.

2.6 Framåtselektion och bakåteliminering

Framåtselektion och bakåteliminering är två algoritmer för att söka bland modeller. Framåtselektion utgår från en modell bestående enbart av intercept. Därefter adderas en förklarande variabel och vi får en ny modell. Variabeln som väljs är den som resulterar i en modell med det lägsta värdet på AIC alternativt BIC. Detta upprepas tills det inte längre går att passa datan något bättre genom att tillföra en ny variabel. Vi har då funnit vår modell. Bakåteliminering fungerar på ett liknande sätt, men istället för att addera förklarande variabler subtraherar vi diton från en modell med alla variabler. Vi är klara när AIC alternativt BIC ej längre kan minska genom att ta bort ytterligare en variabel (Agresti 2013, s. 209–210).

2.7 Residualavvikelser

Residualavvikelse (*residual deviance*) är ett mått på spridningen mellan skattningarna och de faktiska utfallen i en modells träningsdata. För varje enskilt utfall, k, introducerar vi dess avvikelseresidual (*deviance residual*) som

$$r_k^D = \operatorname{sgn}(y_k - \widehat{\pi}_k)\sqrt{d_k},$$

där (Agresti 2013, s. 216)

$$d_k = 2y_k \ln\left(\frac{y_k}{\widehat{\pi}_k}\right) + 2(1 - y_k)\ln\left(\frac{1 - y_k}{1 - \widehat{\pi}_k}\right).$$

Residualavvikelsen är enligt Chambers & Hastie (1992, s. 205) summan av kvadraterna av avvikelseresidualerna, det vill säga $\sum_k (r_k^D)^2$. En låg residualavvikelse innebär en låg spridning mellan skattningarna och utfallen och således passar den bra ihop med datan.

2.8 Hosmer-Lemeshow-testet

Hosmer-Lemeshow-testet (H-L) är ett test för att mäta hur väl en modell passar ihop med träningsdatan. Vi börjar med att skatta $\pi(\mathbf{x}_k)$ för alla n matcher och sorterar dem i storleksordning. Vi delar upp resultatet i g stycken grupper. Enligt standard sätter vi $g = 10$ där varje grupp, i, innehåller n_i observationer. Testet utgår från statistikan

$$\sum_{i=1}^{g} \frac{\left(\sum_j y_{ij} - \sum_j \widehat{\pi}_{ij}\right)^2}{\left(\sum_j \widehat{\pi}_{ij}\right)\left(1 - \left(\sum_j \widehat{\pi}_{ij}\right)/n_i\right)}$$

som är approximativt χ^2-fördelat med $g - 2$ frihetsgrader, där y_{ij} är utfallet för observation j i grupp i (Agresti 2013, ss. 173–174). Vi introducerar nollhypotesen att modellen passar dåligt ihop med datan, som vi förkastar om p-värdet är större än fem procent. Ett högt p-värde innebär inte nödvändigtvis att modellen passar väl ihop med datan, utan enbart att den inte gör det dåligt.

2.9 Sammanblandningsmatris och noggrannhet

När vi predikterar matcher på testdatan finns det fyra utfall vi kan få (se Figur 2). Exempelvis om vi predikterar att matchen kommer sluta i vinst för spelare 1 och så även sker kallar vi prediktionen för sann positiv. Antalet matcher i testdatan klassificerade som sanna positiva betecknas med TP.

		Prediktion	
		1	0
Utfall	1	Sann positiv (TP)	Falsk negativ (FN)
	0	Falsk positiv (FP)	Sann negativ (TN)

Figur 2: Sammanblandningsmatris.

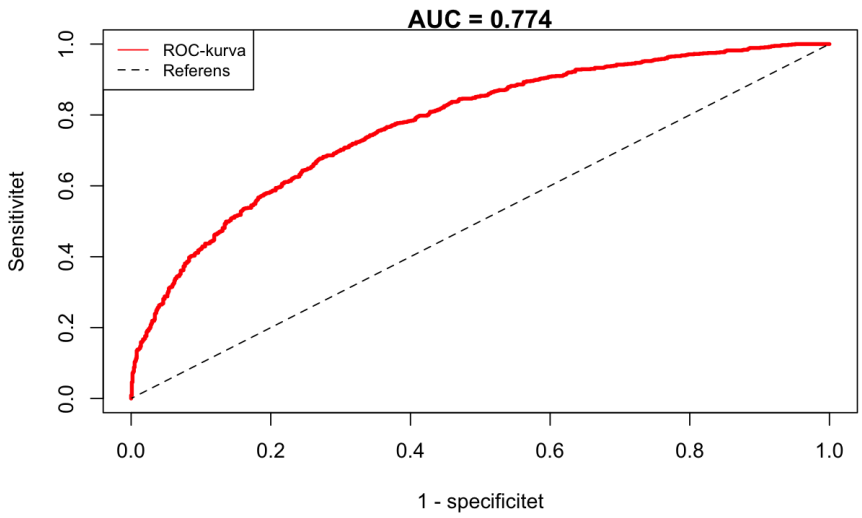
Ur sammanbalndningsmatrisen kan vi få flera olika mått på prediktionsförmågan. En av dem är modellens noggrannhet (*accuracy*) som defineras som (Fawcett 2006)

$$\text{noggrannhet} = \mathbb{P}(\widehat{Y} = Y) = \frac{TP + TN}{TP + TN + FP + FN}.$$

2.10 AUC och ROC-kurva

ROC-kurva (*receiver operating characteristic curve*) är en grafisk illustration av prediktionsförmågan för en modell på testdatan när brytpunkten, b, varieras istället för att bara vara lika med 0.5. Arean under kurvan (AUC) är således ett mått på prediktionsförmågan. Mer specifikt är AUC i detta fall sannolikheten att korrekt kategorisera två godtyckliga matcher där en slutade i vinst och en i förlust. En perfekt modell har AUC $= 1$, medan AUC $= 0.5$ är likvärdigt med att slumpmässigt gissa (Fawcett 2006). Ett exempel på en ROC-kurva syns i Figur 3, där referenslinjen motsvarar godtyckliga gissningar. På y-axeln har vi sensitiviteten och på x-axeln $1 -$ specificiteten, där

$$\text{sensitivitet} = \mathbb{P}(\widehat{Y} = Y \mid Y = 1) = \frac{TP}{TP + FN},$$
$$\text{specificitet} = \mathbb{P}(\widehat{Y} = Y \mid Y = 0) = \frac{TN}{FP + TN}.$$



Figur 3: Exempel på hur en ROC-kurva kan se ut med värde på AUC och godtyckliga gissningar som referens.

3 Data

Datan hämtas från webbsidan CueTracker (2023) som samlar in snookerstatistik från alla professionella matcher, turneringar och spelare. Totalt 383 matcher utgör träningsdatan för modellerna. De är insamlade från nio turneringar av typerna *ranking* (som ger rankingpoäng) samt *invitational* (som ej ger rankingpoäng) från World Snooker Tour för säsongen 2022–23. Mer specifikt handlar det om turneringarna European Masters, British Open, Hong Kong Masters, Northern Ireland Open, Champion of Champions, UK Championship, Scottish Open, English Open och Masters. Efteråt appliceras modellerna på testdatan som består av 94 matcher ur turneringarna Germans Masters och Welsh Open från samma säsong, men som ägde rum efter turneringarna i träningsdatan. Detta för att mäta hur bra de är på att prediktera.

Statistik rörande alla rankade spelare vid början av säsongen samt alla amatörspelare som deltagit i någon av dessa turneringar samlas också in. Totalt rör det sig om 148 spelare, varav 26 är amatörer. Den insamlade datan sammanställs i en databas skriven i `SQL`. Databasen består av flera olika tabeller som sedan kopplas samman. Vilken spelare som får vara spelare 1 väljs slumpmässigt.

3.1 Urval

Ett urval av vilka förklarande variabler som ska undersökas görs. För varje turnering noteras värdland, regerande mästare, typ av turnering, totala prissumman samt vinnarens prissumma (i brittiska pund). För spelarna dokumenteras nationalitet, ranking, kön och tidigare resultat. Slutligen för varje match samlas data in rörande spelare, turnering,

omgång (final, semifinal, et cetera), längd (maximala antalet frames som kan spelas) samt resultat.

Anledningen till att turneringar av typen ranking inkluderas är för att spelarna får rankingpoäng för matcherna. Det gäller däremot inte för invitational, men dessa turneringar är ofta mer prestigefulla och har höga prissummor, därav inkluderingen. I båda fallen är det rimligt att anta att spelarna har stora intressen av att vinna.

Rankingen väljs utifrån hur den såg ut vid slutet av säsongen 2021–22. Rankingen revideras flera gånger under varje säsong vilket innebär att det finns risk att denna data blir missvisande, särskilt för matcherna en bit in i säsongen. Men att ha med den aktuella rankingen vore alldeles för krävande, därav begränsningen.

När det gäller spelarnas tidigare resultat är det som dokumenteras antal vunna respektive förlorade matcher samt vunna respektive förlorade frames i matcher av typerna ranking och invitational från säsongen 2022–23. Anledningen till detta är att aktuell statistik premieras över gammal dito, därav inga tidigare matcher. Men även i likhet med tidigare argumentation inkluderas inga matcher från säsongen 2022–23 på grund av de praktiska och tidskrävande problemen det medför.

De flesta av dessa variabler har även funnits i Li, et al. (2021). Regerande mästare, omgång och kön är tre parametrar som däremot inte var med, men som det torde finnas intresse av att undersöka. Alla variabler och information om dem finns i Tabell 1. Notera att de fem sista observationerna multipliceras med ΔRanking. Detta för att de ej kan stå fristående då de är beroende av motståndaren och ranking är det bästa måttet på skillnaden mellan spelarna. Angående Omgång så består den av fem stycken dummyvariabler. Baslinjen är 32-delsfinal som fås då alla dummyvariabler sätts lika med 0.

Tabell 1: Beskrivning av de förklarande variablerna.

Variabel	Beskrivning	Typ	Nivåer
BättreRanking	Har spelare 1 bättre ranking (enbart Modell 0)	Nominal	1 = Ja 0 = Nej
ΔRanking	Skillnad i ranking	Diskret	$[-122, 120]$
ΔNettoVinster	Skillnad i vinster − förluster	Diskret	$[-56, 49]$
ΔNettoVunnaFrames	Skillnad i vunna − förlorade frames	Diskret	$[-339, 315]$
HemmaSpelare1	Spelas matchen i spelare 1:s hemland	Nominal	1 = Ja 0 = Nej
HemmaSpelare2	Spelas matchen i spelare 2:s hemland	Nominal	1 = Ja 0 = Nej
MästareSpelare1	Är spelare 1 regerande mästare	Nominal	1 = Ja 0 = Nej
MästareSpelare2	Är spelare 2 regerande mästare	Nominal	1 = Ja 0 = Nej
KvinnaSpelare1	Kön spelare 1	Nominal	1 = Kvinna 0 = Man
KvinnaSpelare2	Kön spelare 2	Nominal	1 = Kvinna 0 = Man
Invitational	Typ av turnering	Nominal	1 = Invitational 0 = Ranking
TotalPrissumma	Turneringens totala prissumma	Kontinuerlig	$[315000, 1205000]$
VinnarensAndel	Vinnarens prissumma	Kontinuerlig	$[80000, 250000]$
Frames	Max antal frames som spelas	Diskret	$[7, 19]$
Omgång	Vilken omgång är matchen	Ordinal	32-delsfinal Sextondelsfinal Åttondelsfinal Kvartsfinal Semifinal Final

3.2 Antaganden

Två antaganden i modellarbetet görs. Den första är att matcher som slutar i walkover inte tillför något till modellerna och ignoreras därför. Totalt handlar det om nio matcher i träningsdatan och fyra matcher i testdatan. Motiveringen till detta är att många bettingbolag, som exempelvis DAZN Bet (2023), annullerar alla insatser om en match slutar i walkover. Det andra handlar om hur amatörspelare ska behandlas. Amatörer har nämligen ingen ranking vilket gör att flera variabler skulle sakna värden. Lösningen till

detta är att alla amatörer får en ranking lika med 123 (den sämst rankade spelaren har 122). Således behandlas amatörer som om de vore rankade spelare och de kan inkluderas i modellerna.

3.3 Multikollinearitet

En förutsättning för logistisk regression är en avsaknad av multikollinearitet mellan de förklarande variablerna. Det innebär att en förklarande variabel till en hög grad kan uttryckas med hjälp av andra variabler. Multikollinearitet medför att dessa variablers effekt verkar betydelselös och deras bidrag till modellen försvinner. Att ta bort dessa överflödiga variabler medför även lägre standardfel för de kvarvarande förklarande variablerna (Agresti 2013, s. 208).

3.3.1 Variansinflationsfaktor

Det första måttet vi ska använda är variansinflationsfaktor (VIF). Det är variansinflation som är orsaken till att kollineära variabler kan vara signifikanta när de är tillsammans i en modell, men inte individuellt i samma modell. Dess definition är

$$\mathrm{VIF}_i = \frac{1}{1 - R_i^2},$$

där R_i^2 är förklaringsgraden för hur mycket av variationen i x_i som kan förklaras av de andra x-variablerna (Sundberg 2022, s. 96). Om x_i är ortogonal mot de andra x-variablerna är $\mathrm{VIF}_i = 1$. Kan däremot nästan all variationen förklaras av de andra variablerna, det vill säga $R_i^2 \approx 1$, så medför det att $\mathrm{VIF}_i \to \infty$. Gränsen för accepterade värden sätter vi enligt standard till $\mathrm{VIF}_i < 10$.

För förklarande variabler med fler än 1 frihetsgrad – vilket är sant för Omgång – så används istället generaliserad variansinflationsfaktor (GVIF). Låt oss uttrycka (1) som

$$\mathrm{logit}[\pi(\mathbf{x})] = \alpha + \boldsymbol{\beta}_1^t \mathbf{x}_1 + \boldsymbol{\beta}_2^t \mathbf{x}_2,$$

där $\mathbf{x}_1$ är en 5×1-vektor med alla dummyvariablerna för Omgång (sextondelsfinal, åttondelsfinal, kvartsfinal, semifinal samt final) och $\mathbf{x}_2$ är en kolumnvektor med övriga förklarande variabler. Låt nu $\mathbf{R}_1$ vara korrelationsmatrisen för $\mathbf{x}_1$, $\mathbf{R}_2$ för $\mathbf{x}_2$ och $\mathbf{R}$ för hela $\mathbf{x}$. GVIF kan då enligt Fox (2005) uttryckas som

$$\mathrm{GVIF}_1 = \frac{\det \mathbf{R}_1 \cdot \det \mathbf{R}_2}{\det \mathbf{R}}.$$

Genom att utgå från modellen med alla 14 förklarande variabler utför vi bakåteliminering och tar bort den variabel med högst (G)VIF tills alla har ett värde under 10. Två variabler – TotalPrissumma och ΔNettoVunnaFrames – uppmätte VIF lika med 372.7 respektive 24.04 och tas således bort från framtida modellarbete. Kvarvarande förklarande variabler och deras värden på (G)VIF redovisas i Tabell 2.

Tabell 2: Variansinflationsfaktor för variablerna med låg multikollinearitet.

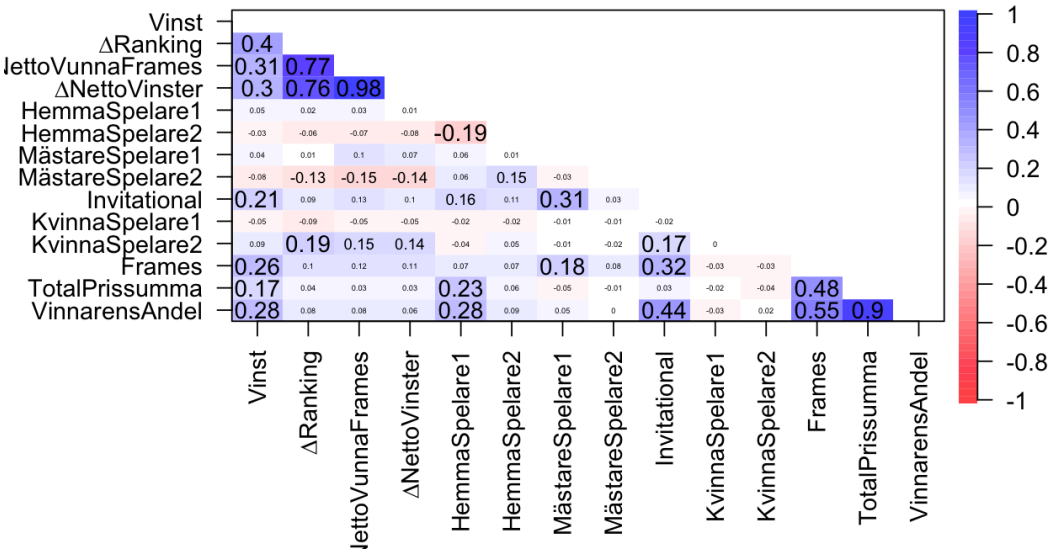
Variabel	(G)VIF	Frihetsgrader
ΔRanking	2.2186	1
ΔNettoVinster	2.229	1
HemmaSpelare1	1.155	1
HemmaSpelare2	1.135	1
MästareSpelare1	1.207	1
MästareSpelare2	1.079	1
KvinnaSpelare1	1.000	1
KvinnaSpelare2	1.000	1
Invitational	1.569	1
VinnarensAndel	2.277	1
Frames	7.037	1
Omgång	7.335	5

3.3.2 Parvis korrelation

Ett annat sätt att mäta multikollinearitet är att undersöka korrelationerna mellan variablerna, det vill säga (Alm & Britton 2008, ss. 245–260)

$$r_{x_i,x_j} = \frac{c_{x_i,x_j}}{s_{x_i}s_{x_j}} = \frac{\sum_k (x_{i,k}-\bar{x}_i)(x_{j,k}-\bar{x}_j)}{\sqrt{\sum_k (x_{i,k}-\bar{x}_i)^2 \sum_k (x_{j,k}-\bar{x}_j)^2}},$$

för alla $i \neq j$. Resultatet av att utföra detta på alla variabler (med undantag för Omgång som ej antar numeriska värden) går att se i Figur 4.



Figur 4: Korrelationsmatris för variablerna.

Franke (2010) beskriver förklarande variabler som ”överdrivet kolineära” ifall deras korrelation överstiger 0.8. Två par uppfyller denna nivå – TotalPrissumma och VinnarensAndel samt ΔNettoVinster och ΔNettoVunnaFrames. Således väljer vi att utesluta en variabel från varje par. I likhet med avsnitt 3.3.1 tar vi bort TotalPrissumma samt ΔNettoVunnaFrames.

4 Metod

Vi börjar med att utgå ifrån den modell som inkluderar alla förklarande variabler – totalt 12 stycken. Vi hittar sedan de modeller som ger de lägsta värdena i AIC, BIC respektive residualavvikelse utfört på träningsdatan. Detta görs med hjälp av en algoritm som undersöker alla möjliga kombinationer ($2^{12} = 4\,096$ stycken) och väljer den med lägst värde i de respektive testen. Detta säkerställer att vi hittar den allra bästa modellen för varje test. För AIC och BIC kommer även framåtselektion samt bakåteliminering att användas. Vi kommer även att jämföra vårt resultat med en modell som enbart består av en binär rankingvariabel. Modellerna genomgår därefter Hosmer-Lemeshow-testet för att säkerställa att de förkastar nollhypotesen om dålig passning av datan. Slutligen undersöks deras respektive prediktiva förmåga utifrån testdatan med hjälp av noggrannhet, sammanblandningsmatriser och AUC.

Hela denna process upprepas även på modeller där vi tillåter tvåvägs-interaktioner mellan variablerna. Bakåteliminering utgår nu ifrån modellen med alla möjliga variabler och tvåvägs-interaktioner. Det är i detta fall ej möjligt att undersöka samtliga möjliga kombinationer ty det skulle kräva resurser som för närvarande ej är tillgängliga. Alla modeller, tester och figurer tas fram i programmeringsspråket `R`.

5 Resultat

Efter att ha granskat över fyratusen kombinationer visar det sig att den modell som minimerar AIC enbart består av ΔRanking. Låt oss kalla den för Modell 1. Vi finner även att Modell 1 minimerar BIC. Modell 1 är också resultatet som ges för framåtselektion samt bakåteliminering med avseende på AIC respektive BIC. Vad gäller residualavvikelse är den modell med lägst värde modellen med alla förklarande variabler, Modell M.

När vi introducerar alla möjliga tvåvägs-interaktioner får vi med framåtselektion även här Modell 1 med avseende på både AIC och BIC. För bakåteliminering på AIC får vi en modell med 20 förklarande variabler, Modell M2A, och på BIC Modell M2B, med lika många variabler. Modell M2A har lägst AIC samt residualavvikelse av alla modeller, medan lägst BIC fortfarande utmäts ur Modell 1.

Vår jämförelsemodell som enbart består av en binär rankingvariabel, Modell 0, skattar sannolikheten att spelare 1 kommer vinna till cirka 66 %, om spelare 1 är bättre rankad, och 32 % annars. Detta oavsett hur stor skillnad i rankingen mellan spelarna är. Det kan jämföras med Modell 1 där vi observerade $\widehat{\pi}(x_{\Delta\text{Ranking}}) \in [0.091, 0.90]$. Värt att notera angående Modell 1 är att den inte alltid predikterar att den bästa spelaren kommer

vinna. Är $\Delta\text{Ranking} \in \{1, 2\}$ så kommer den att presentera den sämre rankade spelaren som den troliga vinnaren.

Alla modellerna klarar Hosmer-Lemeshow-testet med signifikansnivån 5 procent. Angående prediktiv förmåga så har Modell 1 högst AUC och högst noggrannhet av de fem modellerna. (Exakta resultatet går att se i Tabell 3).

Tabell 3: Resultat från testerna utfört på modellerna. Fet text indikerar den bästa modellen för det givna testet.

Modell	AIC	BIC	Res.av.	H-L	AUC	Noggrannhet
0	479.1	486.9	475.1	1.000	0.5889	0.5889
1	457.2	**465.0**	453.2	0.1623	**0.6506**	**0.6000**
M	474.4	541.2	440.4	0.2797	0.6331	0.5889
M2A	**397.6**	551.4	**339.6**	0.6299	0.5541	0.5333
M2B	400.7	541.5	342.7	0.1516	0.5699	0.5667

I Figur 5 kan vi se respektive modells sammanblandningsmatris. Modell 0 och 1 predikterade med ett undantag alla matcher identiskt. Undantaget var en match mellan Luca Brecel och Mark Allen, rankade 12 respektive 14. Modell 0 gav Brecel som den troliga vinnaren medan Modell 1 korrekt predikterde att Allen skulle vinna. Totalt predikterade Modell 1 54 av de 90 matcherna i testdatan korrekt, vilket motsvarar 60 procent. Detta följs tätt av Modeller 0 och M som båda predikterade 53 matcher korrekt. Sämst på att prediktera var de båda interaktionsmodellerna. Modell M2B valde rätt vinnare i 51 matcher och för Modell M2A var det sant för 48 matcher.

		Prediktion	
		1	0
Utfall	1	27	19
	0	18	26

(a) Modell 0.

		Prediktion	
		1	0
Utfall	1	28	19
	0	17	26

(b) Modell 1.

		Prediktion	
		1	0
Utfall	1	28	20
	0	17	25

(c) Modell M.

		Prediktion	
		1	0
Utfall	1	27	24
	0	18	21

(d) Modell M2A.

		Prediktion	
		1	0
Utfall	1	28	22
	0	17	23

(e) Modell M2B.

Figur 5: Sammanblandningsmatriser för modellerna.

Modellernas parametrar och skattade koefficienter samt deras ROC-kurvor går att se i Tabeller 4–8 respektive Figurer 7–11 i bilagan.

6 Slutsats

Då Modell 1 minimerar både AIC och BIC samt Modell M residualvavvikelsen blir det inget stort urval av modeller utan interaktioner att välja mellan. Vad gäller modellerna med tvåvägs-interaktioner presterar de klart bäst i AIC och residualavvikelse, men sämst vad gäller prediktion. Vi noterar särskilt hur ROC-kurvorna för Modeller M2A och M2B (Figurer 10 och 11 i bilagan) delvis ligger under referenslinjen. Detta indikerar att de i vissa fall har sämre prediktionsförmåga än godtyckliga gissningar.

Av de fem modellerna är det två stycken som tillsammans får bäst resultat i alla testerna. (Modell 0 har visserligen högst p-värde i Hosmer-Lemeshow-testet, men det är ej av intresse mer än att den klarar fem procent-gränsen.) Således står valet mellan Modell 1 och Modell M2A som predikterade 60 respektive 53 procent av matcherna korrekt. Modell 1 har bäst resultat i BIC, AUC och noggrannhet. Modell M2A får å andra sidan bäst AIC och residualavvikelse.

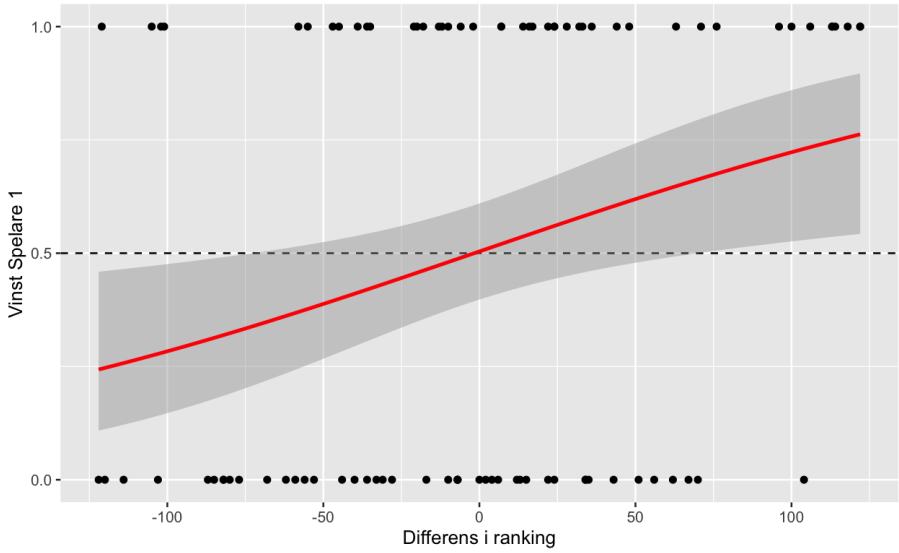
Utifrån detta så är den slutgiltiga modellen vi väljer Modell 1. Skillnaden i att ha en modell med 1 jämfört med 20 förklarande variabler tillsammans med de tester den fick bäst resultat i överväger fördelarna över den mycket mer komplicerade Modell M2A. Att ha en modell enbart beroende av ranking gör modellen förståelig och enkel att använda för personer utan erfarenhet av logistisk regression – satsa (nästan) alltid på den bäst rankade. Den kräver inte heller tillgång till den stora mängd data som Modell M2A gör. Slutligen följer Modell 1 den parismonitetsprincip som presenteras i Sundberg (2022, s. 89) att ”ju enklare modell desto bättre, bara modellen passar tillfredsställande till data.”

7 Diskussion

Det finns flera ytterligare parametrar som skulle vara intressanta att undersöka, men som det ej fanns möjlighet för i detta arbete. En av dem vore att inkludera inbördes möten mellan spelarna, det vill säga resultat från tidigare matcher där spelarna mötts. Detta då det finns exempel på spelare som har extra svårt mot särskilda motståndare jämfört med andra med liknande ranking. Även en enskild modell för varje spelare vore intressant att undersöka i likhet med Willoughby (2002). Vissa spelare kanske har en större fördel av att spela i sitt hemland eller av långa matcher än andra. Sådana variationer kan inte fångas upp i en gemensam modell för 148 olika spelare. Två ytterligare variabler som används i Li, et al. (2021) är ålder samt antal centuries (100 poäng i rad) gjorda. Det är rimligt att anta att seniora spelare har en fördel över juniora diton i en relativt icke-fysisk sport som snooker. Och om en spelare har gjort många centuries kan det ses som en fördel över en motståndare som ej gjort det.

Det skulle även vara att föredra om rankingen hela tiden kunde vara den aktuella och inte från säsongen innan. Det kräver visserligen mer arbete, men det skulle bidra till att kunna bättre beskriva aktuell form på spelarna och därav få säkrare predikteringar. Slutligen vore det intressant att jämföra resultaten med oddsen som ges av olika bettingbolag. Tyvärr har inte denna data varit tillgänglig vid författandet av detta arbete.

I Figur 6 finner vi en graf av den skattade sannolikheten för vinst som en funktion av skillnad i ranking, $\widehat{\pi}(x_{\Delta\text{Ranking}})$, med tillhörande 95-procentigt konfidensintervall. Figuren visar även punkter motsvarande de faktiska utfallen och vi observerar hur det finns flera matcher där den bättre rankade spelaren förlorade (övre, vänstra halvan respektive nedre, högra halvan). Särskilt anmärkningsvärt är där differensen i ranking är stor. I totalt tolv matcher var rankingskillnaden över 50, men den bättre rankade spelaren likväl förlorade. (I fem matcher var skillnaden över 100!) Li, et al. (2021) kallar detta för "chock" när storfavoriten förlorar. Chock beskrivs som en stor utmaning att lyckas prediktera då sannolikheten för det är låg. För att ta ett exempel ur testdatan förlorade världstvåan Judd Trump mot amatörspelaren Daniel Wells (punkten längst upp till vänster i Figur 6). Att korrekt prediktera det utfallet får anses som närmare omöjligt.



Figur 6: Skattad sannolikhet för vinst mot observerade utfall.

En sista reflektion som går att göra från Figur 6 är att konfidensintervallet inkluderar 0.5 i de flesta punkter. Enbart när skillnaden i ranking är större än cirka 75 kan vi göra en signifikant prediktion. Detta kan förklara den – i jämförelse med logistisk regression av andra sporter – låga noggrannheten. Det är dock värt att poängtera att Modell 1 hade en noggrannhet på över 50 procent. Detta är relevant då det innebär att den i alla fall är bättre än att godtyckligt gissa.

Vi kan bevisa att detta resultat även är signifikant. Låt oss introducera nollhypotesen att gissa är minst lika bra som Modell 1. Om Z beskriver utfallet från att godtyckligt gissa alla matcher gäller det att $Z \sim \text{Bin}(90, 0.5)$. Sannolikheten att korrekt gissa minst 54 matcher är

$$\mathbb{P}(Z \geq 54) = 1 - \mathbb{P}(Z \leq 53) = 1 - \sum_{k=0}^{53} \binom{90}{k} 0.5^k (1-0.5)^{90-k} \approx 3.6\ \%.$$

Givet en signifikansnivå på fem procent innebär det att vi förkastar nollhypotesen att ett minst lika bra resultat kunnat ske slumpmässigt. Angående Modeller 0, M, M2A och M2B kan vi inte dra en slutsats gällande ifall de är bättre än att godtyckligt gissa.

Så för att svara på rapportens frågeställning; logistisk regression är till viss del användbar för att prediktera utfallet i snooker. Hosmer & Lemeshow (2000) klassificerar en modell som har AUC $\in [0.7, 0.8[$ som "acceptabel". Ingen av modellerna når upp till den nivån. Dock gäller det att AUC > 0.5 vilket motsvarar att singla slant. Logistisk regression ger oss således möjlighet att prediktera matcher bättre än att gissa, vilket får ses som det allra lägsta som går att begära från en användbar modell. Sen finns det helt klart möjlighet för förbättring.

Referenser

Agresti, A. (2013) *Categorical Data Analysis* (3:e uppl.) Hoboken, NJ: Wiley.

Alm, S. E. & Britton, T. (2008) *Stokastik*. Stockholm: Liber.

Chambers, J. M. & Hastie, T. J., red (1992) *Statistical Models in S*. Pacific Grove, CA: Wadsworth & Brooks/Cole.

CueTracker (2023) *Welcome To CueTracker*. Hämtad 2023-01-21 från `https://cuetracker.net/`.

DAZN Bet (2023) *General Betting Rules*. Hämtad 2023-03-24 från `https://uk.daznbet.com/en/page/general-betting-rules`.

Fawcett, T. (2006) "An introduction to ROC analysis". *Pattern Recognition Letters*, vol. 27, ss. 861–874.

Fox, J. (2005) "Linear Models, Problems". *Encyclopedia of Social Measurement*, vol. 2, ss. 515–522.

Franke, G. R. (2010) "Multicollinearity". *Wiley International Encyclopedia of Marketing*, vol. 2.

Hosmer, D. W. & Lemeshow, S. (2000) *Applied Logistic Regression* (2:a uppl.) New York, NY: Wiley.

Held, L. & Sabanés Bové, D. (2020) *Likelihood and Bayesian Inference* (2:a uppl.) Heidelberg: Springer.

Koo, D. H., et al. (2016) "Logistic Regression of Wins and Losses in Asia League Ice Hockey in the 2014-2015 Season". *International Journal of Performance Analysis in Sport*, vol. 16:3, ss. 871–880.

Li, S., et al. (2021) "Snooker Match Outcome Prediction Using ANN with Inception Structure". *Advances in Intelligent Systems and Computing*, vol. 1398. Cham: Springer, ss. 351–359.

Sundberg, R. (2022) *Lineära Statistiska Modeller*. Stockholm: Stockholms universitet.

Wikimedia Commons (2021) *File:Snooker table drawing 2.svg*. Hämtad 2023-01-21 från `https://commons.wikimedia.org/w/index.php?title=File:Snooker_table_drawing_2.svg&oldid=591135798`.

Willoughby, K. A. (2002) "Winning Games in Canadian Football: A Logistic Regression Analysis". *The College Mathematics Journal*, vol. 33:3, ss. 215–220.

World Professional Billiards and Snooker Association (2022) *Official Rules of the Games of Snooker and English Billiards*. Hämtad 2023-01-21 från `https://wpbsa.com/wp-content/uploads/Rulebook-Website-Updated-May-2022-2.pdf`.

Bilagor

Tabell 4: Skattningar av variablerna i Modell 0.

Variabel	Skattning	Std.fel	p-värde
(Intercept)	-0.7436	0.1595	-3.130e-6***
BättreRanking	1.406	0.2200	1.653e-10***

Tabell 5: Skattningar av variablerna i Modell 1.

Variabel	Skattning	Std.fel	p-värde
(Intercept)	-0.03716	0.1130	0.7423
ΔRanking	0.01854	2.648e-3	2.527e-12***

Tabell 6: Skattningar av variablerna i Modell M.

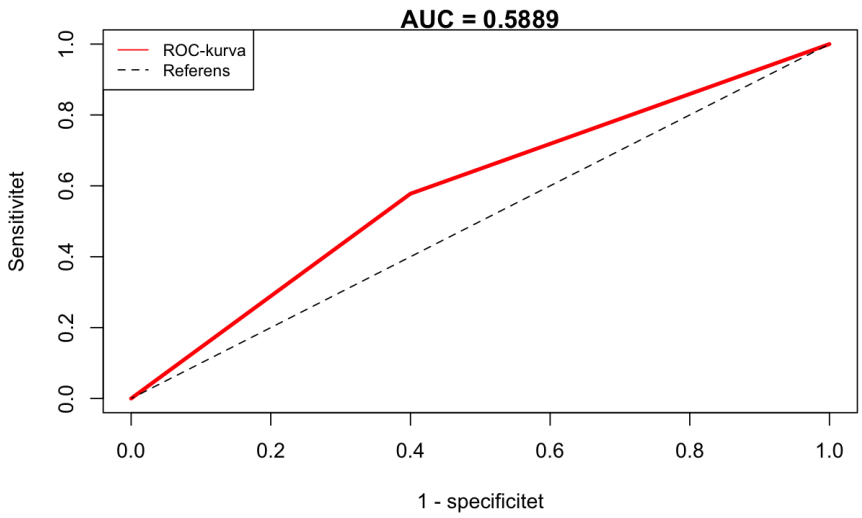
Variabel	Skattning	Std.fel	p-värde
(Intercept)	-0.06758	0.1436	0.6379
ΔRanking	0.03023	0.02159	0.1615
ΔNettoVinster	-2.372e-4	8.983e-3	0.9789
HemmaSpelare1	0.2248	0.3157	0.4765
HemmaSpelare2	0.03908	0.3364	0.9075
MästareSpelare1	0.7081	0.8836	0.4229
MästareSpelare2	-0.3380	0.6454	0.6004
KvinnaSpelare1	-13.52	1455	0.9926
KvinnaSpelare2	15.25	727.9	0.9833
ΔRanking:Invitational	-0.01747	0.01517	0.2495
ΔRanking:VinnarensAndel	5.260e-8	1.025-7	0.6078
ΔRanking:Frames	-1.584e-3	3.386e-3	0.6398
ΔRanking:OmgångFinal	0.1087	0.09845	0.2695
ΔRanking:OmgångSemifinal	-7.986e-3	0.02358	0.7349
ΔRanking:OmgångKvartsfinal	-0.01103	0.01288	0.3918
ΔRanking:OmgångÅttondelsfinal	-3.127e-4	0.01132	0.9780
ΔRanking:OmgångSextondelsfinal	-0.01069	6.505e-3	0.1003

Tabell 7: Skattningar av variablerna i Modell M2A

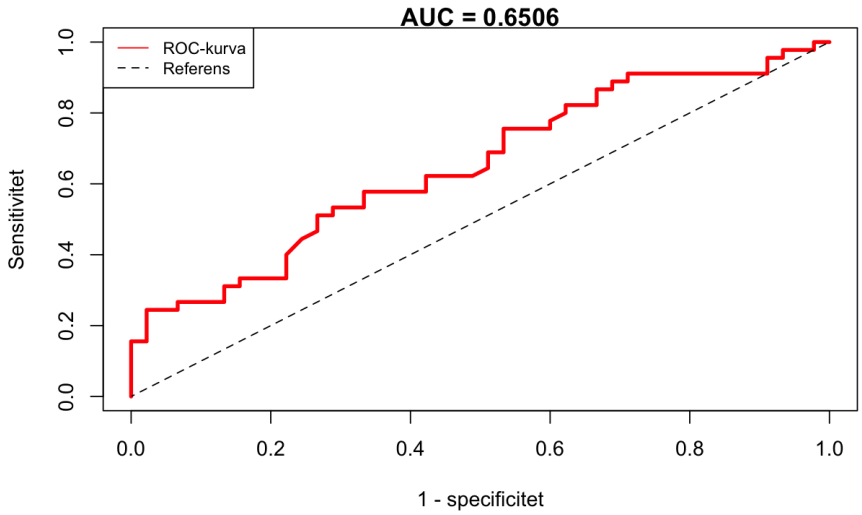
Variabel	Skattning	Std.fel	p-värde
(Intercept)	0.05678	0.1650	0.7307
ΔRanking	-0.1812	0.1183	0.1256
ΔNettoVinster	-0.01274	0.01052	0.2260
HemmaSpelare1	-0.5429	0.4210	0.1971
MästareSpelare1	-0.2824	1854	0.9999
MästareSpelare2	-39.56	2660	0.9881
ΔRanking:NettoVinster	-0.01409	2.831e-3	6.383e-7***
ΔRanking:HemmaSpelare1	0.02789	0.01485	0.06040
ΔRanking:MästareSpelare1	17.06	1854	0.9927
ΔRanking:MästareSpelare2	2.213	151.8	0.9884
ΔRanking:Invitational	-0.5292	0.1775	2.862e-3**
ΔRanking:VinnarensAndel	2.410e-6	1.193e-6	0.04326*
ΔRanking:Frames	0.02152	0.01290	0.09512
ΔRanking:OmgångFinal	0.2099	0.1230	0.08783
ΔRanking:OmgångSemifinal	0.1439	0.07417	0.05231
ΔRanking:OmgångKvartsfinal	9.280e-3	0.03138	0.7674
ΔRanking:OmgångÅttondelsfinal	-0.01647	0.01802	0.3606
ΔRanking:OmgångSextondelsfinal	-0.01331	9.133e-3	0.1450
HemmaSpelare1:MästareSpelare2	150.2	9779	0.9877
ΔRanking:Invitational:NettoVinster	6.799e-3	2.729e-3	0.01273*
ΔRanking:Frames:NettoVinster	1.912e-3	3.993e-4	1.679e-6***
ΔRanking:OmgångFinal:NettoVinster	4.182e-3	0.01294	0.7466
ΔRanking:OmgångSemifinal:NettoVinster	-0.01440	3.729e-3	1.131e-4***
ΔRanking:OmgångKvartsfinal:NettoVinster	-3.847e-3	1.191e-3	1.237e-3**
ΔRanking:OmgångÅttondelsfinal:NettoVinster	-2.666e-3	9.167e-4	3.634e-3**
ΔRanking:OmgångSextondelsfinal:NettoVinster	-2.871e-4	4.111e-4	0.4849
ΔRanking:MästareSpelare2:Invitational	16.52	1172	0.9888
ΔRanking:Invitational:VinnarensAndel	2.653e-6	8.694e-7	2.281e-3**
ΔRanking:VinnarensAndel:Frames	-2.438e-7	1.149e-7	0.03390*

Tabell 8: Skattningar av variablerna i Modell M2B.

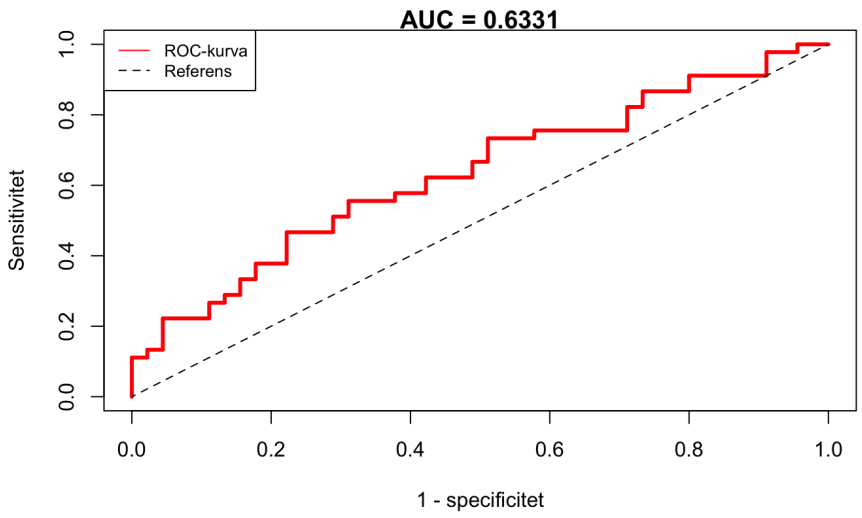
Variabel	Skattning	Std.fel	p-värde
(Intercept)	-0.07281	0.1781	0.6827
ΔRanking	-0.2231	0.1177	0.05798
ΔNettoVinster	-8.747e-3	0.01020	0.3912
HemmaSpelare1	-0.1763	0.3695	0.6333
HemmaSpelare2	0.4517	0.4070	0.2670
MästareSpelare1	-0.3312	3041	0.9999
MästareSpelare2	0.3156	923.3	0.9997
ΔRanking:NettoVinster	-0.01358	2.791e-3	1.154e-6***
ΔRanking:MästareSpelare1	18.15	3041	0.9952
ΔRanking:MästareSpelare2	5.160	307.8	0.9866
ΔRanking:Invitational	-0.5139	0.1785	3.995e-3**
ΔRanking:VinnarensAndel	2.970e-6	1.189e-6	0.01249*
ΔRanking:Frames	0.02480	0.01284	0.05345
ΔRanking:OmgångFinal	0.2356	0.1223	0.05395
ΔRanking:OmgångSemifinal	0.1432	0.07507	0.05650
ΔRanking:OmgångKvartsfinal	0.01658	0.03128	0.5960
ΔRanking:OmgångÅttondelsfinal	-0.01931	0.01788	0.2803
ΔRanking:OmgångSextondelsfinal	-0.01277	8.804e-3	0.1470
HemmaSpelare1:MästareSpelare2	210.7	1.919e4	0.9912
HemmaSpelare2:MästareSpelare2	-72.91	8822	0.9934
ΔRanking:Invitational:NettoVinster	6.447e-3	2.736e-3	0.01845*
ΔRanking:Frames:NettoVinster	1.851e-3	3.946e-4	2.717e-6***
ΔRanking:OmgångFinal:NettoVinster	8.025e-3	1.296e-2	0.5359
ΔRanking:OmgångSemifinal:NettoVinster	-0.01427	3.722e-3	1.260e-4***
ΔRanking:OmgångKvartsfinal:NettoVinster	-3.990e-3	1.174e-3	6.760e-4***
ΔRanking:OmgångÅttondelsfinal:NettoVinster	-2.572e-3	9.042e-4	4.454e-3**
ΔRanking:OmgångSextondelsfinal:NettoVinster	-1.976e-4	3.820e-4	0.6050
ΔRanking:Invitational:VinnarensAndel	2.550e-6	8.709e-7	3.405e-3**
ΔRanking:VinnarensAndel:Frames	-2.897e-7	1.150e-7	0.01177*



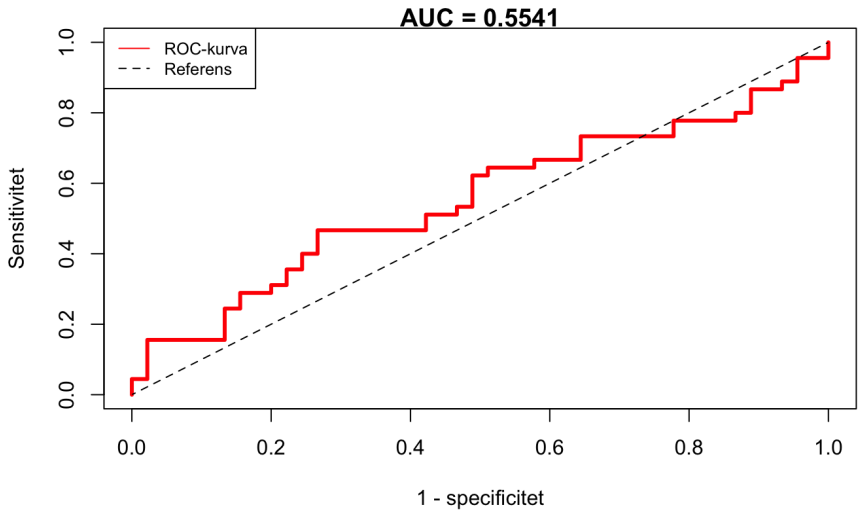
Figur 7: ROC-kurva för Modell 0.



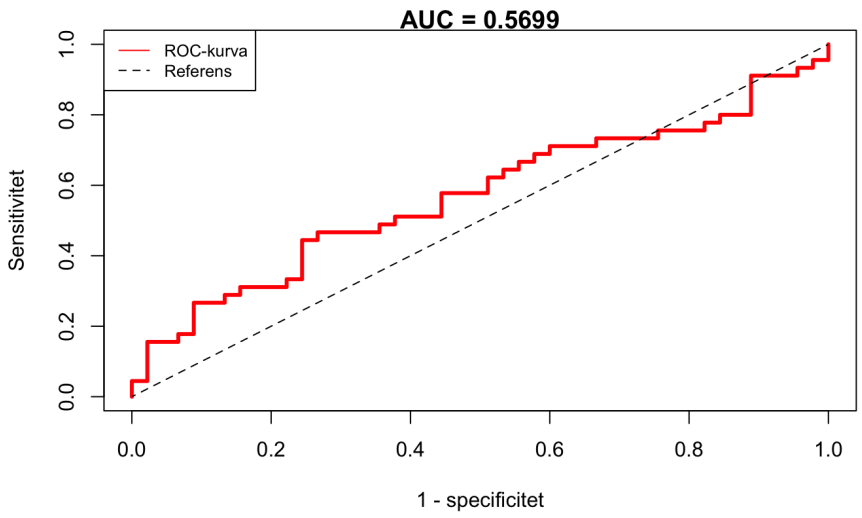
Figur 8: ROC-kurva för Modell 1.



Figur 9: ROC-kurva för Modell M.



Figur 10: ROC-kurva för Modell M2A.



Figur 11: ROC-kurva för Modell M2B.