LUNDS UNIVERSITET
Ekonomihögskolan

Prediktioner av allsvenska fotbollsmatcher

Modellering och testande av strategier för satsning av kapital på spelbolagsmarknaden

Statistiska institutionen

STAH11 HT2018

Kandidatuppsats i statistik (15 HP)

Författare: André Christiansen

Handledare: Björn Holmquist

Abstract

This paper investigates if there is a quantitative relationship between Expected Goals (xG) and results in football games for the Swedish top-ranked league, Allsvenskan. Multicategory logit models are fitted with different types of explanatory xG-variables to see if these models can give better predictions than the bookmakers, and if there's a possibility of generating a positive return using different betting strategies. An indication of a quantitative relationship was found, and a positive return was generated. However, the models couldn't give better predictions in comparison with the bookmakers.

Keywords: Expected Goals, xG, predictions, logistic regression, football, soccer

Sammanfattning

Denna uppsats redogör för om det finns ett kvantitativt samband mellan Expected Goals (xG) och matchresultat för fotbollsmatcher i Sveriges högst rankade fotbollsserie, Allsvenskan. Modeller baserade på multikategorisk logistisk regression anpassas med olika varianter av förklarande xG-variabler för att kunna se om dessa modeller kan ge bättre prediktioner än spelbolagen och generera en positiv avkastning på spelmarknaden med hjälp av spelstrategier. Det fanns en antydan till ett kvantitativt samband mellan xG och matchresultat och positiv avkastning kunde genereras. Modellerna kunde dock inte ge bättre prediktioner vid jämförelser med spelbolagen.

Nyckelord: Expected Goals, xG, prediktioner, logistisk regression, fotboll

Innehåll

1. Inledning 5
2. Begrepp, datainsamling och valda variabler 6
2.1 Expected Goals (xG) 6
2.2 Odds 7
2.3 Datamaterialet 8
2.4 Beskrivning av valda variabler 9
3. Metod och beräkningar 11
3.1 Modellutformning 11
3.2 Spelstrategier 13
4. Resultat 14
4.1 Framställning av modeller 15
4.2 Strategiskt spelande 17
5. Diskussion 18
Referenser 20

1. Inledning

Sedan 80-talet har användandet av dataanalys inom fotbollen ökat kraftigt. De allra första pionjärerna inom området gjorde fotbollsstatistiska fynd tidigare, men det var först då som användandet av datorprogram för att analysera och utvärdera fotbollsspelare blev mer frekvent förekommande. Vid mitten av 00-talet gick det att se att fler och fler fotbollsklubbar anställde statistiker och matchanalytiker för att effektivisera sina organisationer (Kuper, 2011). Företag som Prozone, Wyscout, Opta och STATSports (för att nämna några) har vuxit fram och deras tjänster faktureras idag hos nästan alla professionella fotbollsklubbar i England (Ingle, 2015). Tillgången till positionsdata, där fotbollsspelarnas rörelser registreras med hjälp av GPS-västar, gör att det nu finns betydligt fler parametrar att titta på utöver hur många mål och målgivande passningar som en spelare presterar över en säsong. De professionella fotbollsklubbarna har insett att när de är ute efter nya fotbollsspelare till sitt lag, blir dataanalysen viktig för att undvika dyra felrekryteringar och hitta marginella konkurrensfördelar (MacInnes, 2017).

Det finns idag två fotbollsklubbar som uttalat arbetar med statistiska modeller i alla beslutsled i sin organisation. Ägaren bakom de båda klubbarna är Matthew Benham, en före detta aktiemäklare som insåg att de matematiska kunskaper han hade kunde tillämpas inom fotbollen (Sumpter, 2016). Hans matematiska modeller för att kunna förutse matchresultat, och drivandet av de egna företagen Smartodds och Matchbook, möjliggjorde att han kunde köpa Brentford FC (idag i engelska andraligan) samt FC Midtjylland i danska högstaligan. Där fortsatte han att tillämpa sina matematiska modeller, vilket för FC Midtjyllands del innebar att de under samma säsong som Benham köpte klubben vann danska ligan för första gången någonsin (Biermann, 2015).

Även i Sverige börjar utvecklingen med dataanalys att ta mer plats bland aktörerna på marknaden. I januari 2018 tecknade Bonnier Broadcasting (TV4 och C More), som sänder matcherna i vår inhemska fotbollsliga Allsvenskan, ett avtal med Football Analytics Sweden som bland annat innebär att de får tillgång till Football Analytics Swedens stora databas av match- och spelarstatistik (Bonnier Broadcasting, 2018). För den allmänna fotbollspubliken innebär avtalet främst att de presenterades för termen och prestationsmåttet ”Expected Goals” (”xG”), då detta mått började visas som komplement till den traditionella matchstatistiken såsom antal skott, bollinnehav och antal passningar under den allsvenska säsongen 2018.

Måttet sägs bättre beskriva den faktiska prestationen i matchen. Detta möjliggör att man på ett kvalitativt sätt kan använda måttet för att analysera fotbollslags prestationer över tid, exempelvis se när ett lag över- eller underpresterar (Gregory, 2017). Hur det kvantitativa sambandet mellan xG och den faktiska prestationen i matchen ser ut är dock svårt att hitta något dokumenterat om. Med xG-måttet som grund ska jag därför försöka svara på om det går att kvantifiera det påstådda sambandet mellan xG, matchprestation och slutresultat i matcher genom statistiska modeller, som framställs med hjälp av programmeringsmjukvaran R. Detta görs utefter data tillhandahållen från Football Analytics Sweden (2018) kombinerad med annan insamlade data från football-data.co.uk (Football-Data, 2018).

Statistiska modeller som förutser slutresultat i fotbollsmatcher är ingen ny företeelse (se exempelvis Dixon & Coles, 1997 och Forrest *et al.*, 2005). Oftast används modellerna för att försöka hitta avvikelser i spelbolagens oddssättning och nå en positiv avkastning. Dock har en kvantifierad modell baserad på xG inte gjorts i någon större utsträckning tidigare och det är därför intressant att se hur effektiv en sådan modell är jämfört med spelbolagsmarknaden. Därför ska jag också försöka svara på om de statistiska modeller jag framställer kan prediktera utfallet i fotbollsmatcher bättre än spelbolagen, och om det finns någon strategi för att generera positiv avkastning med hjälp av modellerna. I kommande avsnitt beskrivs de data som analysen baseras på samt de variabler som kommer att användas i de statistiska modellerna. Avsnitt 3 beskriver min valda metod och spelstrategierna. I avsnitt 4 presenteras de empiriska resultaten innan uppsatsen avslutas med en diskussion kring hela arbetsprocessen.

2. Begrepp, datainsamling och valda variabler

Under denna rubrik förklaras Expected Goals och vad odds definieras som. Datamaterialet presenteras tillsammans med de variabler som ska inkluderas i modellerna.

2.1 Expected Goals (xG)

Varje skott i en fotbollsmatch innebär också en chans att göra mål. Varje målchans inkluderar flertalet parametrar, där de viktigaste är positionen på planen (vinkel och distans till mål), teknik (huvud, höger fot eller vänster fot) och spelets skede, exempelvis avslut efter inlägg eller efter genomförd dribbling. Genom att samla in skottdata från ett stort antal matcher, i Football Analytics Swedens fall cirka 200 000 skott, kan ett värde estimeras på hur sannolikt det är att en viss målchans resulterar i ett mål utifrån de historiska data de samlat in. Denna estimerade sannolikhet kallas för Expected Goals (Lidmark Eriksson, 2018).

Måttet kan användas som komplement till traditionell matchstatistik för att visa på vilket lag som faktiskt kom till de bästa målchanserna och därmed presterat bäst. Till exempel, matchen mellan Gefle och Sundsvall den 20 maj 2015 slutade 3-1 till Gefle. De samlade xG-värdena från matchen visade dock 1.78-1.62 till Sundsvall. Detta kan tolkas som att Sundsvall hade ett litet övertag vad det gällde den samlade kvaliteten på målchanser och borde inte ha förlorat matchen. Genom att följa hur många xG ett lag presterar över tid och jämföra mot hur många faktiska mål som görs, går det också att kvalitativt bedöma om ett lag över- eller underpresterar sett till prestationen i xG. Över en säsong brukar skillnaden mellan xG och faktiska mål jämnas ut så att antalet mål och xG sammanfaller med en liten felmarginal. Detta visar bland annat hemsidan *understat.com* för flertalet fotbollsserier. Med Expected Goals ges därför en möjlighet till att förstå varför vissa fotbollsmatcher slutar som de gör, och det finns möjlighet att göra prediktioner utifrån måttet.

2.2 Odds

Inom statistikens sannolikhetslära definieras odds som kvoten mellan sannolikheten att händelse A inträffar och att händelse A inte inträffar. När spelbolagen presenterar sina odds inkluderar denna siffra dock återbetalning av insatsen (Körner & Wahlgren, 2015). Presentationen av odds skiljer sig också mellan olika länder. I denna uppsats är det fokus på de europeiska oddsen, odds i decimalform. Den europeiska formen ger en enkel indikation om vad en satsning på ett potentiellt utfall i en viss match ger i vinst, förutsatt att satsningen är lyckad. Om oddset är 1.7 för ett utfall i en viss match indikerar detta att för en satsad krona ges 70 öre i vinst om utfallet faktiskt inträffar (Sumpter, 2016).

Spelbolagen inkluderar även en vinstmarginal i oddsen de erbjuder. De inkluderar vinstmarginalen eftersom de genom att erbjuda odds som inte är helt sanna skapar sig möjligheter att öka sin egen vinst (Buchdahl, 2017). Genom att titta på genomsnittet av de odds som erbjöds på matchen mellan Malmö FF och Halmstad BK i Allsvenskan den 29 april 2015 kan detta förstås ytterligare. Oddset för hemmavinst var 1.22 (O_H), för oavgjort 6.07 (O_D) och för bortavinst 12.58 (O_A). Om marknaden hade varit rättvis, hade dessa odds konverterade till sannolikheter summerat till 1. Summan Z ger oss,

$$Z = \sum_{i=H}^{A} \frac{1}{O_i} = 1.0639 \qquad i = H, D, A$$

vilket innebär att spelbolagen för denna match i snitt lagt till en marginal på 6.39% på sina erbjudna odds, och att dessa odds i sin tur inte är de ”faktiska” sannolikheterna för de olika

utfallen. För att få fram ”faktiska” sannolikheter behöver de erbjudna oddsen korrigeras genom att multiplicera varje enskilt utfall med Z. Beräkningarna sammanfattas i tabell 1.

Tabell 1. Tabellen beskriver hur spelbolagens marginal räknas ut (Z = 1.0639) och hur denna marginal används för att korrigera sannolikheterna så de summerar till 1. Slh = sannolikhet.

	Hemmavinst	Oavgjort	Bortavinst	
Odds (Spelbolag)	*1.22*	*6.07*	*12.58*	
Slh (Spelbolag)	*0.8197*	*0.1647*	*0.0795*	*= 1.0639*
Odds ("Faktiska")	*1.30*	*6.46*	*13.38*	
Slh ("Faktiska")	*0.7692*	*0.1548*	*0.0747*	*≈ 1*

På en, för spelbolagen, ideal spelmarknad satsas kapital som är proportionerliga mot oddsen de erbjuder, vilket gör att de går med vinst oavsett utfall i matchen (i detta exempel 6.39 kronor per satsad hundralapp).

I praktiken är det dock inte så spelmarknaden agerar och spelbolagen kan välja att räkna om oddsen enligt trenderna inför matchen. Om oddsen till en början visar sig vara felsatta i en match och bolagen får en hög andel satsat kapital på ett visst utfall, kan de välja att korrigera oddsen. Detta för att få spelarna att avstå från att satsa sitt kapital på det utfallet och istället välja ett annat utfall, alternativt avstå helt från att satsa. Bolagen kan också var och en välja att korrigera oddsen och erbjuda mer attraktiva odds än sina konkurrenter för en viss match och utfall. Detta för att locka fler spelare till att satsa sitt kapital i just deras spelbolag eller för att de tror att spelmarknaden och deras egna spelare har fel. Dessa åtgärder görs för att försöka minska risktagandet eller försöka maximera vinsten (Buchdahl, 2017). Oddsen i denna uppsats är dock baserad på marknadens genomsnittliga odds, vilket bör innebära att eventuella avvikelser i den formen beskriven ovan jämnas ut (Nyberg, 2014).

2.3 Datamaterialet

Modellerna byggs utifrån data insamlade från Sveriges högst rankade fotbollsserie, Allsvenskan. Data är insamlad under perioden april 2015 till oktober 2018 och huvuddelen av all data är tillhandahållen av Football Analytics Sweden. Dessa data har kombinerats med data hämtad från *football-data.co.uk.* Avsikten är att modellerna ska bygga på sådan information som finns tillgänglig innan matchen spelas, då det känns i enlighet med vad xG-måttet beskrivs som. Om xG beskriver den faktiska prestationen i matchen, så skulle det innebära att en estimering av hur prestationen kommer vara i kommande match bäst bevisas av hur laget presterat enligt xG i matcherna innan den matchen som ska predikteras. Av den anledningen har de två första omgångarna i varje säsong plockats bort ur datamaterialet. Detta

då de flesta av lagen har spelat en hemmamatch och en bortamatch efter två omgångar, och det ger i sin tur en möjlighet att en eventuell variation mellan prestation på hemmaplan jämfört med bortaplan tas med i beräkningarna.

Reduceringen av datamaterialet innebär att värden från matcherna innan ligger till grund för prediktionerna i nästkommande match. Det kombinerade och reducerade datamaterialet innefattar summerade xG-värden från 850 matcher, gjorda mål i varje enskild match, matchernas slutresultat samt de genomsnittliga odds som spelbolagen presenterar för varje match. I framtagandet av modellerna har nya variabler, baserade på tillhandahållna data, skapats i Excel innan importen till R för vidare analys gjorts. Av dessa 850 matcher är modellerna baserade på matcher spelade från 12 april 2015 till 12 maj 2018, totalt 707 stycken. Resterande 143 matcher utgör alltså de matcher som modellerna kommer prediktera.

2.4 Beskrivning av valda variabler

xGDiff90

Sumpter (2017) har tidigare testat en variabel han kallade för ”xGDiff” i en enkel logistisk regression för att förutspå sannolikheten för en bortavinst. Denna variabel beräknades som

$$xGDiff = \frac{(xG\ Hemmalag + xGA\ \ Bortalag - xGA\ Hemmalag - xG\ Bortalag)}{2}$$

där ”xGA” står för ”Expected Goal Against” och är det antal xG som ett lag ”släpper in” under en match. Detta kan lättare förstås genom att återgå till exempelmatchen Gefle mot Sundsvall den 20 maj 2015, där de samlade xG-värdena från matchen uppgick till 1.78-1.62 i Sundsvalls favör. Här är 1.78 de antal xG som Sundsvall ”gjort” och 1.62 de antal xG som Sundsvall ”släpper in”, alltså deras xGA för just den matchen.

De värden Sumpter inkluderade i de olika delarna i parentesen för beräkning av ”xGDiff”, utgjordes av medelvärden över de 5 senast spelade matcherna. Jag väljer att inkludera denna variabel på ett liknande sätt i mina modeller, där xG och xGA i variabeluträkningen tas från lagens snitt inför varje match (en match = 90 minuter). Detta innebär att inför match nummer två är snittet xG och xGA från föregående match, inför match nummer tre snittet av xG och xGA från de två föregående matcherna och så vidare. Min variabel ”xGDiff90” definieras alltså som

$$xGDiff90 = \frac{(xG90\ Hemmalag + xGA90\ \ Bortalag - xGA90\ Hemmalag - xG90\ Bortalag)}{2}$$

Tanken är att denna variabel ska förklara prestationsförhållandet mellan lagen som möts, där ett positivt värde indikerar att hemmalaget är det bättre laget och ett negativt värde indikerar att bortalaget är det bättre laget.

HxGRatio och AxGRatio

Pseudonymen 11tegen11 (2015) har i en artikel undersökt vad som korrelerar mest med framtida gjorda mål och framtida antal poäng i matcher. I artikeln kommer författaren fram till att det är en ”xGRatio” enligt

$$xGRatio = \frac{Agg\ xG}{(Agg\ xG + Agg\ xGA)}$$

där ”Agg xG” och ”Agg xGA” är det samlade antalet xG och xGA som laget har i matcherna som spelats innan matchen som ska predikteras. Denna kvot tillämpar jag analogt med författaren, med en kvot för hemmalaget och en för bortalaget.

$$HxGRatio = \frac{Agg\ xG\ Hemmalag}{(Agg\ xG\ Hemmalag + Agg\ xGA\ Hemmalag)}$$

$$AxGRatio = \frac{Agg\ xG\ Bortalag}{(Agg\ xG\ Bortalag + Agg\ xGA\ Bortalag)}$$

Dessa två variabler skall alltså ses som ett mått på hur sannolikt det är att laget tar poäng i kommande match, där ett högre värde innebär en större möjlighet för att laget ska ta poäng.

HomeOUOff, AwayOUOff, HomeOUDef och AwayOUDef

Med xG och faktiskt målproduktion går det att kvalitativt avgöra om ett lag för tillfället över- eller underpresterar. Över en säsong brukar skillnaden mellan xG och faktiska mål jämnas ut så att antalet mål och xG sammanfaller med en liten felmarginal. Jag vill med variablerna ”HomeOUOff”, ”AwayOUOff”, ”HomeOUDef” och ”AwayOUDef” försöka kvantifiera över- och underprestationerna både för lagens offensiv och deras defensiv. Dessa variabler skapas enligt

$$HomeOUOff = G90\ Hemmalag - xG90\ Hemmalag$$

$$AwayOUOff = G90\ Bortalag - xG90\ Bortalag$$

$$HomeOUDef = GA90\ Hemmalag - xGA90\ Hemmalag$$

$$AwayOUDef = GA90\ Bortalag - xGA90\ Bortalag$$

där ”G90” och ”GA90” är antal mål per match (en match = 90 minuter) gjorda och insläppta och xG och xGA är detsamma som i variabeln ”xGDiff90”.

LogOdds

Sumpter (2017) föreslår att vid byggandet av en prediktionsmodell för resultat i fotboll bör spelbolagens odds inkluderas. Genom att tillföra en logaritmerad oddskvot i en logistisk regression med andra relevanta variabler, kan prediktionsmodellen hitta felbedömningar och olikheter som kan utnyttjas för att ge möjlighet till högre avkastning. Jag skapar därför variabeln ”LogOdds” vilken beräknas som

$$LogOdds = \ln\left(\frac{1}{(AvgHTr - 1)}\right)$$

Där ”AvgHTr” står för det genomsnittliga oddset för hemmavinst, korrigerat för den vinstmarginal som spelbolagen har på sina odds.

3. Metod och beräkningar

I detta avsnitt ges grunden till analysen och resultatet. Multikategorisk logistisk regression, som är den metod jag kommer att använda i anpassningen av modellerna, förklaras och exempel på hur de predikterade sannolikheterna räknas ut ges. Här presenteras även strategierna som ska testas mot spelmarknaden.

3.1 Modellutformning

Logistisk regression är vanligt förekommande när man har modeller som anpassas till data där den beroende variabeln är av binär karaktär. En generalisering av denna typ av regressionsmetod kan användas när den beroende variabeln är kategorisk, men har fler än två möjliga utfall (Agresti, 2007). Det finns olika typer av modeller som kan anpassas beroende på om den beroende variabeln är nominal eller ordinal samt om de oberoende variablerna i sig är olika kategorivariabler. Dessa modeller kallas med ett samlingsnamn för multicategory logit models (multikategoriska logistiska modeller). I denna uppsats ska jag modellera sambandet mellan resultatet i fotbollsmatcher och variablerna beskrivna i föregående avsnitt. Metoden som beskrivs här har tidigare testats med ”Resultat” som beroende variabel av Nyberg (2014).

Den beroende variabeln ”Resultat” kan tillhöra tre olika kategorier: hemmavinst (H), oavgjort (D) eller bortavinst (A). ”Resultat” är således en nominell, beroende variabel Y med tre olika kategorier. Om

$$\pi_j(\boldsymbol{x}) = P(Y_i = j \mid \boldsymbol{x})$$

där $\sum_j \pi_j(\boldsymbol{x}) = 1$ och **x** är en vektor innehållandes de valda förklarande variablerna enligt

$$\boldsymbol{x} = (xGDiff90, HxGRatio, AxGRatio, \dots, LogOdds)$$

så är $\{\pi_H, \pi_D, \pi_A\}$ de olika sannolikheterna för att Y faller i respektive kategori. Med n oberoende observationer, blir sannolikhetsfördelningen för antalet utfall av de tre olika typerna multinomial (Agresti, 2002).

Med den typ av variabler mitt datamaterial har är en multicategory logit model av typen baseline-category den som är bäst lämpad vid anpassningen. Det innebär att modellen anpassas med en kategori som referens, lämpligen den kategori som förekommer mest frekvent i datamaterialet. I datamaterialet jag bygger modellerna utifrån förekommer H 322 gånger, D 214 gånger samt A 171 gånger. Modellerna byggs därför med H som referenskategori enligt:

$$\log\left(\frac{\pi_j(\boldsymbol{x})}{\pi_H(\boldsymbol{x})}\right) = \alpha_j + \boldsymbol{\beta}'_j\boldsymbol{x} \qquad j = H, D, A$$

Där α_j är de skattade intercepten för respektive kategori och $\boldsymbol{\beta}'_j$ är en transponerad vektor innehållandes de skattade koefficienterna för respektive kategori. Modellen beskriver effekterna av de förklarande variablerna på den beroende variabeln ”Resultat”, där effekterna varierar beroende på vilken responskategori j som väljs. Ur detta fås två ekvationer att lösa för att bestämma parametrarna för respektive kategori (Agresti, 2002). Med H som referenskategori, blir $\boldsymbol{\beta}'_H$ en nollvektor, medan för D och A bestäms parametrarna ur:

$$\begin{aligned}
\log\left(\frac{\pi_D(\boldsymbol{x})}{\pi_A(\boldsymbol{x})}\right) &= \log\left(\frac{\pi_D(\boldsymbol{x})/\pi_H(\boldsymbol{x})}{\pi_A(\boldsymbol{x})/\pi_H(\boldsymbol{x})}\right) \\
&= \log\left(\frac{\pi_D(\boldsymbol{x})}{\pi_H(\boldsymbol{x})}\right) - \log\left(\frac{\pi_A(\boldsymbol{x})}{\pi_H(\boldsymbol{x})}\right) \\
&= \alpha_D + \boldsymbol{\beta}'_D\boldsymbol{x} - (\alpha_A + \boldsymbol{\beta}'_A\boldsymbol{x}) \\
&= (\alpha_D - \alpha_A) + (\boldsymbol{\beta}'_D - \boldsymbol{\beta}'_A)\boldsymbol{x}
\end{aligned}$$

Parametrarna skattas med ML-metoden och uträkningarna görs i R med programmet ”mlogit”, skapat av Yvés Croissant (2018a).

När modellen är specificerad, kan sannolikheten för ett visst utfall i en match *i* uttryckas som:

$$\pi_j(\boldsymbol{x}_i) = P(Y_i = j \mid \boldsymbol{x}_i) = \frac{e^{\alpha_j + \boldsymbol{\beta}_j' \boldsymbol{x}_i}}{\sum_{k=H}^{A} e^{\alpha_k + \boldsymbol{\beta}_k' \boldsymbol{x}_i}}, \quad j = H, D, A \qquad (1)$$

Nämnaren är alltså detsamma för varje enskild sannolikhet, medan täljaren varierar för varje kategori genom att få sin respektive ekvation i exponenten. Samtliga parametrar i referenskategorin är lika med 0, vilket förklarar termen ”1” i nämnaren. De uträknade sannolikheterna för varje enskild kategori blir således

$$\pi_H(\boldsymbol{x}_i) = P(Y_i = H \mid \boldsymbol{x}_i) = \frac{1}{1 + e^{\alpha_D + \boldsymbol{\beta}_D' \boldsymbol{x}_i} + e^{\alpha_A + \boldsymbol{\beta}_A' \boldsymbol{x}_i}}$$

$$\pi_D(\boldsymbol{x}_i) = P(Y_i = D \mid \boldsymbol{x}_i) = \frac{e^{\alpha_D + \boldsymbol{\beta}_D' \boldsymbol{x}_i}}{1 + e^{\alpha_D + \boldsymbol{\beta}_D' \boldsymbol{x}_i} + e^{\alpha_A + \boldsymbol{\beta}_A' \boldsymbol{x}_i}}$$

$$\pi_A(\boldsymbol{x}_i) = P(Y_i = A \mid \boldsymbol{x}_i) = \frac{e^{\alpha_A + \boldsymbol{\beta}_A' \boldsymbol{x}_i}}{1 + e^{\alpha_D + \boldsymbol{\beta}_D' \boldsymbol{x}_i} + e^{\alpha_A + \boldsymbol{\beta}_A' \boldsymbol{x}_i}}$$

3.2 Spelstrategier

Utifrån de anpassade modellerna ska försök till att generera positiv avkastning på oddsmarknaden göras. Detta ska ske genom att testa tre olika strategier. Den första strategin är mest inkluderad för att svara på de två första frågorna denna uppsats undersöker, det vill säga om det finns ett kvantifierbart samband mellan xG och matchresultat samt om en modell baserat på xG ger bättre prediktioner än spelbolagen. Övriga två strategier försöker utnyttja eventuella avvikelser i spelbolagens oddssättning för att nå en positiv avkastning. Resultatet från det strategiska spelandet räknas ut genom att exportera de sannolikheter som modellen ger från R, importera dem till Excel och sedan jämföra dem mot spelbolagens sannolikheter. I Excel framställs därefter tabeller för att redovisa antalet korrekt predikterade matcher samt avkastningen för varje strategi.

Strategi 1: Högst sannolikhet

Modellernas prediktionsförmåga testas där högsta värdet av sannolikheterna $\{\pi_H, \pi_D, \pi_A\}$ i varje match väljs och satsas på. De testas alltså som att de bedömer sannolikheterna för de olika utfallen helt korrekt, och ingen hänsyn tas till avvikelser i spelbolagens oddssättning.

Vikten läggs vid antalet rätt prediktioner i utvärderingen, men även avkastningen tas med som ett jämförelsevärde för att se om mina modeller hittar vissa avvikelser hos spelbolagen.

Strategi 2: Hitta undervärderade odds

Sumpter (2016) föreslår en enkel regel vid beslut om man ska satsa sitt kapital på ett visst utfall eller inte. Han menar att satsningen ska ske när ens uppskattade sannolikhet för ett utfall är större än spelbolagens presenterade sannolikheter. Översatt till parametrarna i denna uppsats kommer satsning att ske när

$$\pi_j(\boldsymbol{x}_i) > \frac{1}{AvgTr_j} \qquad j = H, D, A$$

där ”AvgTr” är det genomsnittliga, korrigerade oddset för utfall *j* i match *i*. Hittar modellerna fler än en positiv avvikelse mellan modellernas sannolikheter och spelbolagens korrigerade sannolikheter i en viss match kommer inget kapital att satsas, eftersom det inte är tydligt vilket utfall som ska väljas av de som avviker. Viktigast med denna strategi är att avkastningen blir så hög som möjligt och antalet rätt prediktioner blir oväsentligt.

Strategi 3: Satsa på bortasegrar

I datamaterialets 850 matcher förekommer 322 hemmasegrar, 214 oavgjorda matcher och 171 bortasegrar. Andelen bortasegrar är klart lägst, vilket naturligt innebär att sannolikheten för att en bortavinst inträffar också är lägre. Detta återspeglas även i de insamlade oddsen från spelbolagen där snittoddset för en bortaseger är 4.52, jämfört med 4.12 för oavgjort och 2.65 för hemmavinst. I enlighet med strategi 2 så sker satsning när modellerna hittar avvikelser, men här tas ingen hänsyn till om modellerna hittar fler än en avvikelse för en viss match eftersom det är uttalat att det är bortasegrar det ska satsas på. Nås en tillräckligt hög andel av korrekt predikterade bortavinster bör kapitalvinsten ändå överstiga förlusterna då oddsen för bortavinst i genomsnitt är högre. Även här blir avkastningen viktigast i analysen av resultatet.

4. Resultat

I detta avsnitt beskrivs det hur modellerna tagits fram och analyserats utifrån de valda strategierna. Modellernas skattade parametrar redovisas och strategierna testas. Resultatet för modellerna och för de olika strategierna presenteras samt kommenteras.

4.1 Framställning av modeller

Med hjälp av paketet ”mlogit” i R skapas fyra olika modeller baserat på datamaterialet innefattande 707 allsvenska matcher. Utformningen sker genom att tillföra en variabel eller en grupp av variabler från vektorn $\boldsymbol{x}$ i taget. Detta dels för att se om det går att göra bra prediktioner med en så enkel modell som möjligt, dels för att se om de olika typerna av variabler har någon påverkan på det slutgiltiga resultatet och kunna jämföra dem med varandra. Uträkningarna i R för koefficienterna i modellerna 1 och 2 ges i tabell 2 och för modell 3 och 4 i tabell 3.

Tabell 2. Tabell som visar utformningen av modellerna M1 och M2, med parameterskattningar och p-värden. Två skattningar görs till varje modell eftersom det är två modeller som skattas, bortavinst relativt till hemmavinst samt oavgjort relativt till hemmavinst.

M1			**M2**		
Variabel	β_j	p-värde	Variabel	β_j	p-värde
Intercept $_A$	*-0.029*	*0.774*	*Intercept* $_A$	*0.591*	*0.194*
Intercept $_D$	*-0.399*	*<0.001*	*Intercept* $_D$	*0.060*	*0.900*
xGDiff90 $_A$	*-1.439*	*<0.001*	*xGDiff90* $_A$	*0.920*	*0.320*
xGDiff90 $_D$	*-0.623*	*<0.001*	*xGDiff90* $_D$	*0.112*	*0.899*
			HxGRatio $_A$	*-7.291*	*0.005*
			HxGRatio $_D$	*-2.456*	*0.331*
			AxGRatio $_A$	*6.024*	*0.021*
			AxGRatio $_D$	*1.796*	*0.473*

Tabell 3. Tabell som visar utformningen av modellerna M3 och M4, med parameterskattningar och p-värden. Två skattningar görs till varje modell eftersom det är två modeller som skattas, bortavinst relativt till hemmavinst samt oavgjort relativt till hemmavinst.

M3			M4		
Variabel	β_j	p-värde	Variabel	β_j	p-värde
$Intercept_A$	0.561	0.238	$Intercept_A$	-0.267	0.595
$Intercept_D$	-0.025	0.960	$Intercept_D$	0.520	0.312
$xGDiff90_A$	0.801	0.388	$xGDiff90_A$	0.777	0.405
$xGDiff90_D$	-0.114	0.899	$xGDiff90_D$	-0.066	0.942
$HxGRatio_A$	-7.138	0.007	$HxGRatio_A$	-2.456	0.363
$HxGRatio_D$	-2.215	0.382	$HxGRatio_D$	0.940	0.717
$AxGRatio_A$	6.001	0.021	$AxGRatio_A$	1.255	0.640
$AxGRatio_D$	1.519	0.545	$AxGRatio_D$	-1.738	0.500
$HomeOUOff_A$	0.020	0.906	$HomeOUOff_A$	0.238	0.185
$HomeOUOff_D$	-0.097	0.577	$HomeOUOff_D$	0.085	0.635
$HomeOUDef_A$	0.177	0.356	$HomeOUDef_A$	-0.142	0.476
$HomeOUDef_D$	0.483	0.016	$HomeOUDef_D$	0.245	0.228
$AwayOUOff_A$	0.146	0.457	$AwayOUOff_A$	-0.011	0.955
$AwayOUOff_D$	0.337	0.099	$AwayOUOff_D$	0.241	0.235
$AwayOUDef_A$	-0.301	0.110	$AwayOUDef_A$	-0.266	0.175
$AwayOUDef_D$	-0.266	0.170	$AwayOUDef_D$	-0.213	0.276
			$LogOdds_A$	-1.724	<0.001
			$LogOdds_D$	-1.199	<0.001

I programmet R görs även ett LR-test, vilket visar om modellen har minst en signifikant parameter. Samtliga modellers test visar att där är minst en parameter som är signifikant, dock är det noterbart att för M4 är det enbart variabeln ”LogOdds” som är signifikant trots att de enklare modellerna har parametrar som är signifikanta.

Koefficienterna för de skattade parametrarna i modellerna ska tolkas med försiktighet. I en linjär modell kan koefficienterna direkt tolkas som de marginella effekterna som de förklarande variablerna har på den beroende variabeln. I mina multikategoriska modeller är koefficienterna relativa till resultatet H (Croissant, 2018b). Det innebär att koefficienterna kan tolkas som att för en ökning med en enhet av en viss förklarande variabel, förändras sannolikheten för oavgjort eller bortavinst relativt hemmavinst med den givna parameterskattningen, uttryckt i logaritmerade odds. Minskningar och ökningar i sannolikheten för ett visst resultat kan uttryckas genom att ta e^{β_j}, vilket ger den relativa

risken för att resultatet blir det som väljs istället för referensresultatet H. Således ska exempelvis koefficienten för variabeln ”xGDiff90” och resultat A i modell M1, -1.439, tolkas som att för varje enhets ökning i ”xGDiff90” så blir sannolikheten för en bortavinst relativt en hemmavinst med $e^{-1.439} = 0.237$ gånger mindre, givet att allt annat hålls konstant.

4.2 Strategiskt spelande

Datamaterialet med de 143 matcherna som ska predikteras matas in i R för att skatta sannolikheter för de olika resultaten enligt formel 1. Det är dessa sannolikheter som sedan ligger till grund för besluten i de olika strategierna presenterade i föregående avsnitt. Tabellerna 4, 5 och 6 presenterar resultaten av de olika spelstrategierna. I dessa beskrivs antalet rätt och avkastning i både absoluta tal och i procent för att kunna jämföra de olika modellernas resultat. Insatsen är 1 SEK per match oberoende av strategi. Tabellerna 4 och 6 visar också spelbolagens ”modell” som referens (SB i tabellerna), det vill säga vad resultatet hade blivit om strategierna följts enbart utifrån de odds som spelbolagen presenterar. Strategi 2 bygger på att hitta avvikelser, och därför visar tabell 5 inte spelbolagens resultat.

Tabell 4. Resultat för spelande på matcher enligt det utfall som modellen ger högst sannolikhet för. Antal rätt, antal spelade matcher och avkastningen för spelandet visas i både absoluta tal och i procent. SB = Spelbolagen.

Strategi 1: Högst sannolikhet		
Modell	Rätt / Spelade (%)	Avkastn. i SEK (%)
M1	*74/143 (51.7%)*	*-8.98 (-6.28%)*
M2	*75/143 (52.4%)*	*-4.23 (-2.95%)*
M3	*75/143 (52.4%)*	*-5.14 (-3.59%)*
M4	*82/143 (57.3%)*	*3.08 (2.15%)*
SB	*83/143 (58%)*	*5.64 (3.94%)*

Tabell 5. Resultat för spelande på matcher där modellen hittar avvikelser i spelbolagens sannolikhetsbedömningar. Antal rätt, antal spelade matcher och avkastningen för spelandet visas i både absoluta tal och i procent. SB = Spelbolagen.

Strategi 2: Undervärderade odds		
Modell	Rätt / Spelade (%)	Avkastn. i SEK (%)
M1	*29/85 (34.1%)*	*-10.1 (-11.88%)*
M2	*27/85 (31.8%)*	*-17.4 (-20.44%)*
M3	*26/78 (33.3%)*	*-3.36 (-4.31%)*
M4	*47/87 (54%)*	*6.43 (7.39%)*
SB	*	*

Tabell 6. Resultat för spelande på bortasegrar när modellen ger högst sannolikhet för en bortaseger. Antal rätt, antal spelade matcher och avkastningen för spelandet visas i både absoluta tal och i procent. SB = Spelbolagen.

Strategi 3: Bortasegrar		
Modell	Rätt / Spelade (%)	Avkastn. i SEK (%)
M1	*20/35 (57.1%)*	*2.23 (6.37%)*
M2	*23/41 (56.1%)*	*4.6 (11.2%)*
M3	*22/40 (55%)*	*2.52 (6.3%)*
M4	*31/50 (62%)*	*9.91 (19.82%)*
SB	*30/46 (65.2%)*	*11.22 (24.39%)*

Då det viktigaste för strategi 1 var antalet korrekt predikterade matcher är det anmärkningsvärt att endast en marginell förbättring sker vid tillförandet av fler variabler i modellerna M2 och M3 i förhållande till den enklaste modellen (M1). Det är först när variabeln ”LogOdds” läggs till i M4 som antalet korrekt predikterade matcher ger en märkbar ökning. Spel på bortasegrar ger en positiv avkastning och detta gäller för samtliga modeller. Det är dock endast M4 som uppvisar positiv avkastning för samtliga strategier.

5. Diskussion

Denna uppsats redogör för om det finns något kvantifierbart samband mellan xG, matchprestation och slutresultat samt om statistisk modellering av detta eventuella samband kan ge fler antal korrekta prediktioner jämfört med spelbolagen. Det undersöks även om det finns en möjlighet att generera positiv avkastning med statistiska modeller baserade på xG kombinerat med strategiskt spelande.

Med tanke på att den allra enklaste modellen, M1, baserad på variabeln ”xGDiff90” predikterar rätt i drygt hälften av fallen och därmed hamnar nära spelbolagen i antalet rätt, verkar det antyda att det finns ett kvantifierbart samband. Adderandet av ytterligare variabler ger dock ingen märkbar effekt på antalet korrekta prediktioner förrän när variabeln ”LogOdds” läggs till. Detta beror förmodligen på att de andra variablerna är förhållandevis lika varandra, och baseras alla utifrån xG-värden. Då syftet med uppsatsen är att undersöka det eventuella sambandet mellan xG och matchresultat, hade variabler som förklarar andra företeelser i fotbollsmatcher varit ett bra tillskott till modellerna för att öka antalet korrekta prediktioner. Även om det finns indikationer på ett kvantifierbart samband är det inte heller säkert att detta är korrekt modellerat. För framtida studier kan det vara intressant att testa xG i exempelvis en Poissonregression med faktiska mål som beroende variabel för att utifrån det kunna göra prediktioner kring matchresultat.

Modellen M4 är den modell som ger bäst avkastning av de statistiska modeller som framställts. Precis som Sumpter förespråkar är det just tillförandet av ”LogOdds”-variabeln som gör att modellen hittar avvikelser i spelbolagens oddssättning och därför är det också enbart M4 som ger en positiv avkastning (7.39%) när strategin att spela på matcher där oddsen anses undervärderade tillämpas. Modellen överträffar dock inte spelbolagen när det gäller spel på bortasegrar, utan med den strategin hade det varit bäst att spela på bortaseger när spelbolagen presenterar ett lågt odds på det utfallet för att maximera sin avkastning.

Vad gäller avkastningen för modellerna bör dessa tolkas något tvetydigt. Datamaterialet som modellerna bygger på är förhållandevis litet, där enbart 707 matcher används för att bygga modellerna och antalet matcher som predikteras är 143 stycken. Det finns en viss risk att det positiva resultatet bygger på slump, och resultatet hade därför kunnat variera kraftigt från vad som presenteras här om faktiskt satsande hade skett. Exempelvis ges en hög, positiv avkastning för strategin med spel på bortasegrar, men detta är baserat på förhållandevis lite prediktioner. Över en längre tid hade troligen inte avkastningen blivit densamma. Samtidigt är de insamlade oddsen i denna uppsats genomsnittliga, och bedömningsavvikelserna därför små, vilket antyder att avkastningen hade kunnat maximeras genom att hitta de mest spelvärda oddsen. Till framtida studier kan det vara vettigt att inkludera ett statistiskt test i framställandet av resultatet för att se om avkastningen som uppnås beror på slump eller ej.

Resultaten indikerar att det både finns ett kvantifierbart samband mellan xG, matchprestation och slutresultat samt att det finns en möjlighet att generera positiv avkastning. Dock kunde antalet korrekta prediktioner inte överträffa spelbolagen. Med mer data hade modellen kunnat testats på fler matcher, och det kan därför vara intressant att pröva sambanden och resultaten som framställts här när ytterligare säsonger har spelats och mer historiska xG-data finns tillhanda för att säkerställa slutsatserna i denna uppsats.

Referenser

Agresti, A. (2002). Categorical Data Analysis, John Wiley & Sons Inc., New Jersey.

Agresti, A. (2007). An Introduction to Categorical Data Analysis, John Wiley & Sons Inc., New Jersey.

Biermann, C. (2015). Plötsligt i Herning, *Offside #4 2015, s. 84-93*. (Översättning: Stefan Carlsson)

Bonnier Broadcasting (2018). TV4 och C More i samarbete med Football Analytics Sweden, länk: https://www.bonnierbroadcasting.com/senaste-nytt/pressmeddelanden-innehall/2018/tv4-och-c-more-i-samarbete-med-football-analytics-sweden/ (2018-12-14)

Buchdahl, J. (2017). The Wisdom of the Crowd, länk: http://www.football-data.co.uk/The_Wisdom_of_the_Crowd_updated.pdf (2018-12-22)

Croissant, Y. (2018a). Package "mlogit": Multinomial Logit Models (hämtad i R) (2018-10-28)

Croissant, Y. (2018b). Estimation of multinomial logit models in R: The mlogit Packages, länk: https://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf (2018-11-27)

Dixon, M. & Coles, S. (1997). Modelling Association Football Scores Inefficiencies in the Football Betting Market. *Journal of the Royal Statistical Society: Series C (Applied Statistics), No. 2, pp. 265-280*.

Football Analytics Sweden (2018). *"Data från Allsvenskan (xG)"*. Personlig kontakt.

Football-Data (2018). "*Data från Allsvenskan"*, länk: http://www.football-data.co.uk/sweden.php (2018-11-20)

Forrest, D., Goddard, J. & Simmons, R. (2005). Odds-setters as forecasters: The case of English football. *International Journal of Forecasting 21, pp. 551-564.*

Gregory, S. (2017). BLOG: Expected Goals in context, länk: https://www.optasportspro.com/about/optapro-blog/posts/2017/blog-expected-goals-explained/ (2019-01-16)

Ingle, S. (2015). Number crunching: why analytics still adds up at Brentford, länk: https://www.theguardian.com/sport/blog/2015/oct/04/number-crunching-analytics-brentford-football (2018-12-14)

Kuper, S. (2011). Nördarnas revansch, *Offside #6 2011, s. 79-94*. (Översättning: Johan Nilsson)

Körner, S. & Wahlgren, L. (2015). Statistisk dataanalys, Studentlitteratur AB, Lund.

Lidmark Eriksson, O. (2018). Så funkar radars och förväntade mål, länk: https://www.fotbollskanalen.se/fotbollslabbet/sa-funkar-radars-och-forvantade-mal/ (2019-01-07)

MacInnes, P. (2017). Expected Goals and Big Football Data: the statistics revolution that is here to stay, länk: https://www.theguardian.com/football/2017/mar/30/expected-goals-big-football-data-leicester-city-norwich (2018-12-14)

Nyberg, H. (2014). A Multinomial Logit-based Statistical Test of Association Football Betting Market Efficiency. *Discussion Paper No. 380*. University of Helsinki (Center of Economic Research).

Sumpter, D. (2016). Mer än en sport: Fotbollens matematik, Volante, Stockholm. (Översättning: Stefan Lindgren)

Sumpter, D. (2017). Part two: Is there a magical betting formula?, länk: https://www.pinnacle.com/en/betting-articles/Betting-Strategy/part-two-magical-betting-formula/FTN2ELAEUYHGAR34 (2018-12-29)

11tegen11 (2015). The best predictor for future performance is Expected Goals, länk: http://11tegen11.net/2015/01/05/the-best-predictor-for-future-performance-is-expected-goals/ (2018-12-29)