Stockholms universitet

Vinstprediktion för ishockeymatcher i NHL

Milla Esko

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2017:21
Matematisk statistik
Juni 2017

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholms universitet

Matematisk statistik
Stockholms universitet
Kandidatuppsats **2017:21**
http://www.math.su.se/matstat

Vinstprediktion för ishockeymatcher i NHL

Milla Esko*

Juni 2017

Sammanfattning

Syftet med denna uppsats var att hitta en modell för prediktion av NHL-matcher som endast bygger på data från tidigare matcher under aktuell och föregående säsong. Data består av säsongerna 2005/2006-2016/2017. Två olika typer av modeller undersöktes, som hanterade saknade värden för variablerna på olika sätt. För en av modellerna uteslöts de tre första matcherna för varje lag och för den andra användes viktade variabler. Logistisk regression användes för båda modellerna med utfallet för hemmalaget som responsvariabel. Högsta andelen korrekta prediktioner var 59.2%.

*Postadress: Matematisk statistik, Stockholms universitet, 106 91, Sverige.
E-post: milla.esko@gmail.com. Handledare: Tom Britton och Benjamin Allévius.

Abstract

The object of this thesis was to find a model for predicting the outcomes of NHL games, only using data from previous games during the current and previous season. Logistic regression, with the outcome for the home team as the response variable, and data the seasons 2005/2006-2016/2017 were used to fit two types of models. The first model dealt with missing values for the first games of each season by excluding each teams' first three games for every season. The second model used weighted explanatory variables, with data from both the current and previous season in the same variable, to deal with the problem. The results were not much better than chance, and the highest accuracy achieved was 59.2%.

Förord

I denna uppsats presenteras mitt kandidatarbete i matematisk statistik på Stockholms universitet under vårterminen 2017. Jag vill passa på att tacka mina handledare Tom Britton och Benjamin Allévius för all hjälp och stöd under arbetets gång.

Jag har använt mig av programvaran R och den kod jag skrivit för det här arbetet finns tillgänlig på https://github.com/millaesko/kandidat.

Innehåll

1 Inledning

Ishockey är en populär sport i Nordamerika och norra Europa, med NHL (National Hockey League) som den högst rankade ishockeyligan i världen. I varje match möts två lag med sex spelare var (inklusive målvakt) och det laget som har gjort flest mål vinner. Matchen är indelad i tre perioder om 20 minuter (effektiv tid), dvs totalt 60 minuter. Om matchen inte avgjorts inom ordinarie matchtid, försöker man i första hand avgöra matchen i övertid och i sista hand i straffar. I övertid har båda lagen tre spelare på isen samt målvakt. Övertiden varar i fem minuter eller tills något av lagen gör mål och därmed vinner matchen, annars avgörs matchen i straffar.

Grundserien i NHL består av 30 lag, 23 från USA och 7 från Kanada, som är uppdelade i två konferenser, Östra och Västra. Varje konferans är i sin tur uppdelad i olika divisioner, och indelningen i dessa har varierat något genom åren. Normalt spelar varje lag under grundserien 82 matcher, 41 hemma och 41 borta, mellan oktober och april. De bästa 8 lagen i varje konferans går sedan vidare till Stanley Cup.

Till skillnad från andra sporter, såsom basket, fotboll och amerikansk fotboll, tycks det ha varit svårare att hitta bra prediktiva modeller för resultaten i ishockeymatcher. En del av förklaringen kan vara att slumpen spelar en större roll för målgörandet och därmed resultaten i ishockey. En annan kan vara att ishockey är en mindre populär sport och att det därför inte har ägnats lika mycket tid som andra sporter. På senare år har dock flera NHL-lag börjar satsa på den statistiska analysen av spelet, och flera sidor på nätet dedikerar sig åt att samla och analysera matchdata[1].

1.1 Syfte och frågeställning

Syftet med detta arbete är att undersöka om det går att hitta en modell för prediktion av utfall i NHL-matcher, som endast utgår från tidigare prestationer på lagnivå under aktuell och närmast föregående säsong. Idén är att ta fram en modell med variabler som bygger på lättillgänglig data utan användandet av avancerade statistikor såsom t.ex. Corsi, som mäter skillanden i skott på mål då lagen spelar med samma antal spelare.

För att eliminera den kanske mest kända faktorn för framgång, d.v.s. om laget spelar hemma eller borta, fokuserar jag här på vinster i hemmamatcher.

1.2 Tidigare forskning

Flera studier har fokuserat på stokastisk modellering av ishockeymål[2, 3] eller på att finna samband mellan faktorer under spelets gång och resultat[4, 5]. De studier som utgår från tidigare prestationer använder sig ofta av mer avancerade statistikor, t ex på spelarnivå. Weissbock och Inkpen har undersökt olika faktorer som är kända innan matchen och deras prediktiva förmåga i NHL-matcher. Den modell de funnit med bäst prediktionsförmåga predikterade 59.8% av matcherna rätt i ett urval av 517 matcher spelade mallan februari och april 2013.

Weissbock fann även att det tycks finnas en teoretisk övre gräns på 62% för korrekta prediktioner för matcher i NHL[1, 6]. Studien återskapades med samma datamängd av Gianni Pischeda, som lyckades med en av sina modeller prediktera 61.54% av matcherna[7].

2 Data

I det här arbetet har jag använt mig av data från Hockey Reference[8] för säsongerna 2005/2006 till 2016/2017, både från grundserien och slutspelen. Data är i tabellform med ett antal variabler som ges matchvis. De variabler som är av intresse här utgörs av följande:

- Hemmalag.
- Bortalag.
- Resultat: 1 om hemmalaget vinner, 0 om bortalaget vinner.
- Mål för hemmalaget.
- Insläppta mål för hemmalaget.
- Skott för hemmalaget.
- Skott mot hemmalaget.
- Antal utvisningsminuter för hemmalaget respektive bortalaget.

2.1 Variabler

De variabler jag undersökt har beräknats ur ovanstående data. Två variabler för att mäta framgångarna i slutspelen har också inkluderats, antal spelade matcher och antal vinster. Samtliga undersökta variabler finns i Tabell 1 samt Tabell 2 nedan. Variabelnamnen bygger på en del förkortningar som är vanligt förekommande i ishockeysammanhang. De uttryck som används förklaras kortfattat nedan:

- Tm (team), avser hemmalaget.
- Opp (opposing team), avser bortalaget.
- Hth (head to head), avser möten mellan två bestämda lag.
- S (shots), avser hemmalagets skott på mål.
- SA (shots against), avser motståndarens skott på mål.
- Pims (penalties in minutes), utvisningsminuter.
- GF (goals for), mål för hemmalaget.

- GA (goals against), mål mot hemmalaget/insläppta mål.
- WP (win percent), vinstprocent.
- ls (last season), avser föregående säsong. T.ex. WP.ls - vinstprocent föregående säsong.
- Diff (difference), målskillnad. D avser differensen mellan två variabler.
- GP (games played), antal spelade matcher.

Variabel	**Beskrivning**
HthWP	Vinst% i matcher mot samma motståndare, föregående och innevarande säsong.(Hth - head to head)
LastHth	Utfall senaste matchen mot samma motståndare.
DiffHth	Genomsnittlig målskillnad mot samma motståndare, föregående och innevarande säsong.
S.Tm	Genomsnittligt antal skott-skott emot under innevarande säsong, hemmalaget.
S.Opp	Genomsnittligt antal skott-skott emot under innevarande säsong, bortalaget.
Pims.Tm	Genomsnittligt antal utvisningsminuter - motståndares utvisningsminuter under innevarande säsong. Hemmalaget.
Pims.Opp	Genomsnittligt antal utvisningsminuter - motståndares utvisningsminuter under innevarande säsong. Bortalaget.
LastGame.Tm	Utfall senaste match hemmalaget.
LastGame.Opp	Utfall senaste match bortalaget.
LastGame2.Tm	Utfall förrförra matchen, hemmalaget.
LastGame2.Opp	Utfall förrförra matchen, bortalaget.
Last3Games.Tm	Utfall senaste tre matcherna, hemmalaget.
Last3Games.Opp	Utfall senaste tre matcherna, bortalaget.
PlayoffWins.Tm	Antal vunna matcher i slutspelen föregående säsong, hemmalaget.
PlayoffWins.Opp	Antal vunna matcher i slutspelen föregående säsong, bortalaget.
PlayoffGP.Tm	Antal spelade matcher i slutspelen föregående säsong, hemmalaget.
PlayoffGP.Opp	Antal spelade matcher i slutspelen föregående säsong, bortalaget.
OvertimeP.Tm	Andel matcher som avgjordes i övertid eller straffar föregående säsong, hemmalaget.
OvertimeP.Opp	Andel matcher som avgjordes i övertid eller straffar föregående säsong, bortalaget.
GFLastGame.Tm	(Goals for last game), gjorda mål föregående match, hemmalaget.

GALastGame.Tm	(Goals against last game), insläppta mål föregående match, hemmalaget.
GFLastGame.Opp	(Goals for last game), gjorda mål föregående match, bortalaget.
GALastGame.Opp	(Goals against last game), insläppta mål föregående match, bortalaget.
GF.Tm	Genomsnittlig antal mål per match, innevarande säsong.
GF.Opp	Genomsnittlig antal mål per match, innevarande säsong.
GA.Tm	Genomsnittlig antal insläppta mål per match, innevarande säsong.
GA.Opp	Genomsnittlig antal insläppta mål per match, innevarande säsong.
WP.Tm	Vinst% under aktuell säsong.
WP.Opp	Vinst% under aktuell säsong.
GA.Tm.ls	Genomsnittligt antal insläppta mål per match, föregående säsong.
GA.Opp.ls	Genomsnittligt antal insläppta mål per match, föregående säsong.
GF.Tm.ls	Genomsnittlig antal mål per match, föregående säsong.
GF.Opp.ls	Genomsnittlig antal mål per match, föregående säsong.
WP.Tm.ls	Vinst% under föregående säsong.
WP.Opp.ls	Vinst% under föregående säsong.
GFD	GF.Tm-GF.Opp.
GAD	GA.Tm-GA.Opp.
SD	S.Tm - S.Opp.
PimsD	Pims.Tm-Pims.Opp.
HWP	Vinst% i hemmamatcher för hemmalaget under föregående säsong.
RWP	Vinst% i bortamatcher för bortalaget under föregående säsong.
WPD	WP.Tm - WP.Opp.
GPD	PlayoffGP.Tm - PlayoffGP.Opp.
WPD.ls	WP.ls - WP2.ls.
OvtD	OvertimeP.Tm - OvertimeP.Opp.
LG	LastGame.Tm - LastGame.Opp.
LG3	Last3Games.Tm - Last3Games.Opp.
PWD	PlayoffWins.Tm-PlayoffWins.Opp
RegWin	Vinst% matcher som avgörs i ordinarie matchtid, föregående säsong.
OTWin	Vinst% matcher som avgörs i övertid (overtime), föregående säsong.
SOWin	Vinst% matcher som avgörs i straffar (shoot out), föregående säsong.

Tabell 1: Variabler som undersöktes i Modell 1.

Variabel	Definition
DiffHth	genomsnittlig målskillnad i matcher mot samma lag
HthWP	Vinst% i matcher mot samma lag
LastHth	Senaste match mot samma lag
PlayoffGP.Tm	Antal matcher spelade i sluspelen förra året, hemmalag
PlayoffGP.Opp	Antal matcher spelade i sluspelen förra året, bortalag
WinP.Opp	Vinst% för bortalag
WinP.Tm	Vinst% för hemmalag
GA.Tm	genomsnittligt antal insläppta mål, hemmalag
GF.Tm	genomsnittligt antal gjorda mål, hemmalag
GF.Opp	genomsnittligt antal gjorda mål,bortalag
GA.Opp	genomsnittligt antal insläppta mål, bortalag
S.Tm	Genomsnittlig Skott% per match
S.Opp	Genomsnittlig Skott% per match
Pims.Tm	Genomsnittlig antal utvisningsminuter per match, hemmalag
Pims.Opp	Genomsnittlig antal utvisningsminuter per match, bortalag

Tabell 2: Förklarande variabler som undersöktes i Modell 2.

2.2 Träningsmängd, valideringsmängd och testmängd

Det som är av intresse här är att anpassa en modell med så bra prediktionsförmåga som möjligt för utfallet av NHL-matcher. Eftersom NHL-matcher alltid slutar med att hemmalaget antingen vinner eller förlorar, har vi alltså en binär och kategorisk responsvariabel. I sådana fall brukar man använda sig av logistisk regression, som beskrivs kortfattat i Avsnitt 3.2

För att unvika problem med överanpassning av modellen till data har jag delat in den i en träningsmängd, en valideringsmängd och en testmängd. Träningsmängden består av matchdata från säsongerna 2006/2007 till 2014/2015, valideringsmängden av säsongen 2015/2016 och testmängden av säsongen 2016/2017.

Data från säsong 2005/2006 samt från slutspelen har bara använts för beräkning av variablerna rörande prestationer för föregående säsong. Modellen har anpassats på träningsmängden och dess prediktiva förmåga har kontinuerligen testats mot valideringsmängden. Den slutgiltiga modellen har därmed valts efter hur väl den predikterat matchutfallen på valideringsmängden och dess prediktionsförmåga testas slutligen på testmängden. De metoder jag använt för att utvärdera modellens prediktionsförmåga beskrivs kortfattat i Avsnitt 3.3.

2.3 Hantering av saknade värden

En del av dessa variabler medför saknade värden för säsongernas första matcher för varje lag. Jag har undersökt två alternativ för att komma runt problemet.

1. Exkludera de första matcherna för varje lag.
2. Vikta variablerna med lämpligt $\lambda \in [0, 1]$. Om exempelvis v_n är vinstprocenten för aktuell säsong och v_f vinstprocenten för föregående säsong, använder vi istället en variabel v för vinstprocent, som ges av

$$v = v_n \cdot \lambda + v_f \cdot (1 - \lambda).$$

I Modell 1 har jag exkluderat de 3 första matcherna för varje lag. I Modell 2 har jag testat att vikta en del av variablerna så som beskrivits ovan för olika värden för λ. Modellerna beskrivs närmare i Avsnitt 4.1 respektive Avsnitt 4.2

2.4 Inledande dataanalys

Under säsongerna 2006/2007 till 2014/2015 (d.v.s. den period som utgör träningsmängden) avgjordes 76.3% av matcherna i ordinarie matchtid, 10.1% i övertid och 13.6% i straffar. Hemmalaget vann 54.7% av matcherna under perioden och 56.1% av matcherna som avgjordes under ordinarie matchtid. Av de matcher som avgjordes i övertid eller straffar stod hemmalaget för 50.3% av vinsterna och bara 47.4% av dem som avgjordes i straffar.

Variabel	**Hemmalag**	**Bortalag**
Mål	2.95	2.66
Skott	30.80	28.91
Utvisningsminuter	11.73	12.45

Tabell 3: Mål, skott och utvisningsminuter under säsongerna 2006/2007 - 2014/2015.

I Tabell 3 ser vi fördelningen genomsnittliga antalet mål, skott och utvisningsminuter för hemma- respektive bortalaget under samma period. Vi ser att hemmalaget i genomsnitt har färre utvisningsminuter, skjuter fler skott och gör fler mål än bortalaget. Detta kan tyda på att hemmalaget spelar bättre än bortalaget, att det finns en tendens att döma till hemmalagets fördel vad gäller utvisningsminuter eller att skotten räknas fel. De händelser som troligen påverkas i mindre utsträckning av subjektiva bedömningar vid datainsamling är gjorda mål och därmed också resultat. Bedömningen av händelser så som skott på mål och handlingar som leder till utvisningar är däremot mer subjektiva, vilket ökar risken för felaktigheter [9, 10].

Detta är värt att ha i åtanke, men eventuella felaktigheter blir svåra att kontrollera, särskilt för den senasate säsongen som spelas under samma tidsperiod som denna uppsats skrivs. Möjliga tillkortakommanden i datainsamlingen bortses därför ifrån i den fortsatta analysen.

3 Teori

I detta kapitel redogörs kortfattat för de statistiska metoder som jag använt mig av för att anpassa och utvärdera modellerna.

3.1 Generaliserade linjära modeller

I enkla linjära regressionsmodeller ges sambandet mellan responsvariabeln Y_i och förklarande variabeln X av

$$Y_i = \alpha + \beta X_i.$$

Vidare antas att $Y_i \sim N(\mu_i, \sigma^2)$, och att alltså

$$\mu_i = E(Y_i) = \alpha + \beta x_i. \tag{1}$$

I generaliserade linjära modeller antas istället att fördelningen för Y_i tillhör familjen exponentialfördelningar. Gemensamt för dessa fördelningar är att täthetsfunktionen kan skrivas på formen

$$f(x;\theta) = h(x)e^{\theta x - A(\theta)}.$$

Till skillnad från i fallet med linjära regressionsmodeller ges förhållandet mellan μ_i och förklarande variabeln av en *länkfunktion*, $X\beta = g(\mu_i)$[11]. Därutöver är modellen linjär i parametrarna β_i, inte nödvändigtvis i förklarande variablerna X_i.

3.2 Logistisk regression

Varje NHL-match slutar i antingen vinst eller förlust. Om vi sätter, som responsvariabel, utfallet från hemmalagets perspektiv får vi, för $i = 1, \ldots, n$,

$$Y_i = \begin{cases} 1 & \text{om hemmalaget vinner;} \\ 0 & \text{om bortalaget vinner.} \end{cases}$$

Med andra ord så är $Y_i \sim Bernoulli(p_i)$, där $p_i = p(X_i)$, beror på de förklarande variablerna X_i. Bernoullifördelningen, som är ett specialfall av binomialfördelningen, tillhör exponentialfamiljen och därför kan generaliserade linjära modeller användas. Eftersom $E(Y_i) = p_i$, så får vi enligt (1)

$$p_i = \boldsymbol{X}_i\boldsymbol{\beta}.$$

Denna modell är dock problematisk, då $p_i \in [0, 1]$ medan $\boldsymbol{X}_i\boldsymbol{\beta}$ inte är bundet av samma intervall. Detta problem kommer man runt genom att som länkfunktion använda *logit*-funktionen,

$$g(p_i) = \text{logit}(p_i) = \log \frac{p_i}{1 - p_i}.$$

Sambandet mellan responsvariabeln $p_i = P(Y_i = 1)$ och förklarande variabler $\boldsymbol{X}$ ges då av

$$g(E(Y_i)) = \log \frac{p_i}{1 - p_i} = \boldsymbol{X}\boldsymbol{\beta} = \beta_0 + \beta_1 X_1 + \ldots + \beta_n X_n.$$

Detta innebär att när exempelvis X_1 ökas från k till $k + 1$, så ökar oddsen $\frac{p}{1-p}$ med faktor e^{β_1} [12].

3.3 Utvärdering av prediktiv förmåga

Regressionsmodellen som jag använt mig av anpassar sannolikheten för att hemmalaget ska vinna en match, givet värden för de förklararande variablerna. Modellen utvärderas sedan genom att på samma sätt beräkna sannolikheterna för hemmavinst för matcherna i valideringsmängden givet de förklarande variabler som använts i regressionsmodellen. Eftersom denna utvärdering görs efter varje förändring i modellen, är valet av metod för utvärdering av stor betydelse för utvecklingen av slutmodellen. Det finns många olika metoder för att utvärdera prediktiva förmågan hos logistiska regressionsmodeller. Här presenteras några metoder som jag undersökt när jag valt modell.

3.3.1 Log loss

Log loss-funktionen ges av

$$\text{log loss} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{M} y_{ij} \log(p_{ij}),$$

eller då det (som i detta fall) rör sig om en binär responsvariabel av

$$\text{log loss} = -\frac{1}{N} \sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i).$$

y_i utgörs av de faktiska resultaten som ska predikteras, och p_i av modellens anpassade sannolikheter för vinst. Log loss fungerar genom att straffa felaktiga prediktioner med hög säkerhet hårt. En perfekt modell skulle ha log loss lika med noll[13].

Fördelen med log loss är att den tar hänsyn till osäkerheten i prediktionerna och hur mycket den skiljer sig från den predikterade kategorin. En nackdel är att det saknas en övre gräns, vilket gör resultatet mer svårtolkat än mått som R^2 i linjär regression.

Som ett jämförelsevärde kan vi tänka oss föjande situation. Om sannolikheten för vinst i varje match sattes till 0.547, d.v.s. andelen matcher som vanns av hemmalaget under perioden 2006/2007-2014/2015, skulle vi få log loss = 0.691 när vi testar prediktionen mot valideringsmängden.

3.3.2 Klassifikationsförmåga

Som beskrivits ovan, anpassar modellen olika sannolikheter till de olika matcherna baserat på värderna för de förklarande variablerna. Dessa anpassade sannolikheter beräknas sedan för matcherna i valideringsmängden och testmängden. Om vi säger att varje sådan sannolikhet som är större än 0.5 har predikterat en vinst, kan vi räkna ut klassifikationsförmågan som antalet korrekta prediktioner. Skulle vi exempelvis ha samma sannolikheter om i stycket ovan, d.v.s. $p_i = 0.547, i = 1, \ldots, 1230$, predikteras alltså att hemmalaget vinner varje match. För valideringsmängden skulle vi då ha klassificierat 52.9% av matcherna rätt.

3.3.3 AUC

AUC står för Area under curve, och syftar på arean under den s.k. ROC-kurvan (receiver operating characteristic curve). ROC-kurvan plottar sambandet mellan andel sanna positiva prediktioner och falska positiva prediktioner med olika brytpunkter. Idén är att brytpunkten 0.5 som normalt används inte nödvändigtvis är optimal. I många sammanhang anses AUC över 0.8 vara bra, medan AUC runt 0.5 anses motsvara slumpen. Denna metod har dock blivit ifrågasatt [14] och jag har valt att inte använda AUC som beslutsgrundande, men inkluderat värdena för de slutliga modellerna.

4 Modellval

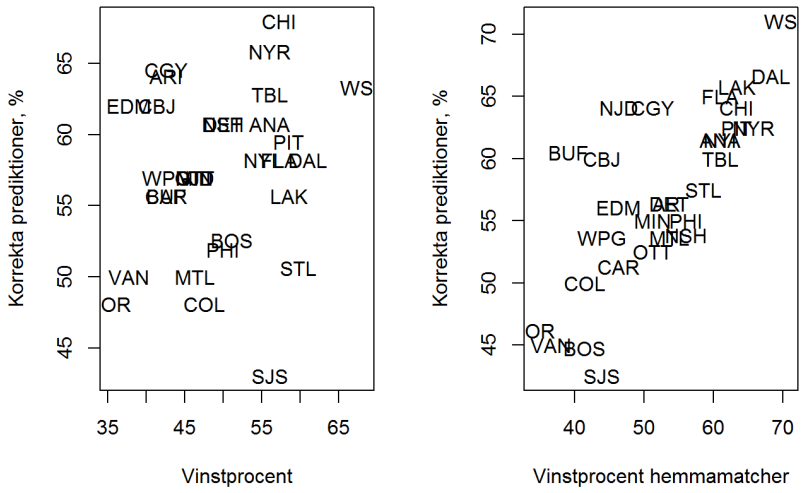
I början har jag inkluderat samtliga ursprungliga förklarande variabler i modellen. Nästa steg var att ta bort eller modifiera en variabel i taget och jämföra den nya modellen med avseende på log loss, och behålla de ändringar som minimerade detta. Modifikationerna i varje steg bestod exempelvis av att ta differensen av två variabler, så som vinstprocent för hemmalag och bortalag, eller transformation av förklarande variabler. Varje transformation motiverades helt av att öka modellens prediktionsförmåga, eftersom en tolkning av modellen är av sekundärt intresse.

Modellerna anpassades i R med hjälp av funktionen glm från paketet stats. Koefficientskattningarna har gjorts enligt ML-metoden[15].

4.1 Modell 1

De variabler jag undersökte här beskrivs i Tabell 1 (Avsnitt 2.1). Jag testade att ta bort de 3,5 och 10 första matcherna för varje lag under varje säsong, och fann att ju fler matcher jag exkluderade desto sämre resultat fick jag. Därför valde jag att endast exkludera de 3 första matcherna för varje lag.

Den slutliga modellen innehåller de förklarande variablerna i Tabell 4 och i Tabell 5 finns de värden på klassifikationsförmåga, log loss och AUC för modellen när den undersöktes mot valideringsmängden.



Figur 1: Andel korrekta prediktioner i hemmamatcher mot vinstprocent i samtliga matcher till vänster och hemmamatcher till höger för de 30 NHL-lagen under säsong 2015/2016.

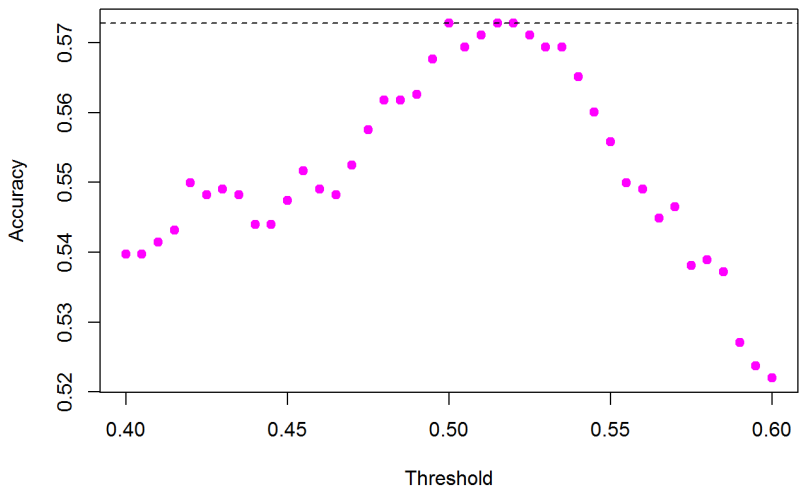
Variabel	$\hat{\beta}$	**Standarfel**	**z-värde**	**p-värde**
Intercept	-0.5433	$3.052e-01$	-1.780	0.075058
sqrt(HthWP)	-0.09170	$8.510e-02$	-1.078	0.281205
LastHth	0.03267	$4.865e-02$	0.672	0.501879
SD	0.03309	$4.240e-03$	7.806	$5.91e-15$
exp(Pims.Tm)	$-2.210e-05$	$8.640e-05$	-0.256	0.798156
Pims.Opp	-0.03439	$1.354e-02$	-2.539	0.011113
LastGame.Opp	-0.01565	$4.149e-02$	-0.377	0.705971
exp(LastGame2.Opp)	0.04723	$2.417e-02$	1.954	0.050646
Last3Games.Tm	-0.01169	$2.574e-02$	-0.454	0.649852
PlayoffWins.Tm	0.007952	$5.728e-03$	1.388	0.165054
sqrt(OvertimeP.Tm)	0.3730	$1.843e-01$	2.024	0.043021
exp(GFLastGame.Tm)	$-2.779e-05$	$3.901e-05$	-0.712	0.476240
exp(GALastGame.Tm)	$-4.460e-05$	$3.752e-05$	-1.189	0.234547
exp(GF.Opp)	-0.007431	$2.380e-03$	-3.123	0.001793
GA.Opp	0.1218	$5.366e-02$	2.270	0.023192
exp(WP.Tm)	0.3781	$1.172e-01$	3.225	0.001260
exp(GA2.ls)	0.01252	$3.531e-03$	3.547	0.000389
exp(GF2.ls)	-0.006050	$3.550e-03$	-1.704	0.088380
log(WP.ls)	0.4643	$1.361e-01$	3.411	0.000648
log(WP.ls):exp(GF.ls)	0.007800	$7.270e-03$	1.073	0.283328

Tabell 4: Förklarande variabler i slutliga modellen, skattade koefficienter, standardfel och p-värde.

I Figur 1 är andelen korrekta prediktioner plottade mot vinstprocent för säsong 2015/2016 för varje lag. Det lag med högst andel korrekta klassifikationer, 68%, är Chicago Blackhawks (CHI) och det med lägst, 43%, är San Jose Sharks (SJS). I bilden till höger syns ett tydligare linjärt samband mellan vinstprocent och andel korrekta prediktioner än i den till höger, där de lag med högre andel vunna hemmamatcher också haft högre andel korrekta prediktioner. San Jose Sharks har också här lägst andel korrekta prediktioner (42.5%) medan Washington Capitals (WSH) har högst med 72.1%.

Mängd	**Klassifikationsförmåga**	**Log Loss**	**AUC**
Hela säsongen	57.3 %	0.679	0.585
Ordinarie- och Övertid	56.9%	0.679	0.589
Ordinarie matchtid	58.2%	0.671	0.609

Tabell 5: Modell 1: Resultat på valideringsmängden.



Figur 2: Andel korrekta prediktioner för olika värden för brytpunkten.

Av de undersökta matcherna i valideringsmängden blev 57.3 % korrekt klassificierade. Totalt predikterade modellen att 75.1% av matcherna skulle sluta i vinst, medan bara 53.4% faktiskt gjorde det. Av vinsterna klassificerades 19.7% felaktigen som förluster och av förlusterna klassificerades 69.2% felaktigen som vinster. De lag med hög andel vinster får troligen därför högre andel korrekt predikterade matcher. I Figur 2 ser vi andelen korrekta prediktioner om brytpunkten för prediktion av vinst varieras (se Avsnitt 3.3.2). Ett annat värde för brytpunkten hade alltså inte ökat andelen korrekta prediktioner för valideringsmängden. Eftersom modellen inte har särskilt bra prediktionsförmåga, undersökte jag om denna blev bättre genom att minska på antalet säsonger i

träningsmängden och därmed få olika skattningar för koefficienterna. I Tabell 6 finns värden på log loss, AUC och klassifikationsförmåga för varje försök.

Den datamängd som gav högst andel korrekta prediktioner (57.9%) och lägst log loss (0.678) på valideringsmängden bestod av säsongerna 2009/2010-2014/2015. I Tabell 7 finns skattade koefficienterna, standardfel, z-värde och p-värde för modellen med reducerade träningsmängden.

Valideringsmängd	**Klassifikationsförmåga**	**Log Loss**	**AUC**
2007/2008-2014/2015	57.2 %	0.6785	0.5849
2008/2009-2014/2015	57.5 %	0.6784	0.5851
2009/2010-2014/2015	57.9%	0.6783	0.5838
2010/2011-2014/2015	57.3 %	0.6784	0.5842
2011/2012-2014/2015	57.8%	0.6785	0.5844
2012/2013-2014/2015	56.7%	0.6792	0.5831
2013/2014-2014/2015	57.2%	0.6789	0.5823
2014/2015	57.3%	0.6832	0.5742

Tabell 6: Modell 1: Resultat på valideringsmängder för reducerad träningsmängd.

Variabel	$\hat{\boldsymbol{\beta}}$	**Standardfel**	**z-värde**	**p-värde**
Intercept	-0.4781	$3.977e-01$	-1.202	0.229237
sqrt(HthWP)	-0.1519	$1.058e-01$	-1.436	0.151143
LastHth	0.03769	$6.109e-02$	0.617	0.537241
SD	0.03889	$5.331e-03$	7.295	$2.98e-13$
exp(Pims.Tm)	$-8.699e-05$	$1.313e-04$	-0.662	0.507739
Pims.Opp	-0.04620	$1.757e-02$	-2.629	0.008552
LastGame.Opp	-0.01701	$5.164e-02$	-0.329	0.741841
exp(LastGame2.Opp)	0.07267	$3.011e-02$	2.414	0.015791
Last3Games.Tm	0.01333	$3.206e-02$	0.416	0.677486
PlayoffWins.Tm	0.01010	$7.255e-03$	1.392	0.164012
sqrt(OvertimeP.Tm)	0.2967	$2.384e-01$	1.244	0.213326
exp(GFLastGame.Tm)	$-5.156e-05$	$6.455e-05$	-0.799	0.424426
exp(GALastGame.Tm)	$-5.743e-05$	$7.206e-05$	-0.797	0.425452
exp(GF.Opp)	-0.007257	$3.180e-03$	-2.282	0.022461
GA.Opp	0.1222	$6.890e-02$	1.773	0.076203
exp(WP.Tm)	0.2663	$1.428e-01$	1.864	0.062256
exp(GA2.ls)	$0.01672e$	$5.011e-03$	3.337	0.000848
exp(GF2.ls)	$-0.008150e$	$4.915e-03$	-1.658	0.097294
log(WP.ls)	0.5368	$1.751e-01$	3.066	0.002172
log(WP.ls):exp(GF.ls)	-0.002061	$1.040e-02$	-0.198	0.842961

Tabell 7: Förklarande variabler i slutliga modellen, skattade koefficienter, standardfel och p-värde för modellen med reducerade träningsmängden (2009/2010-2014/2015).

4.2 Modell 2

De variabler som jag undersökte i den här beskrivs i Tabell 2 i Avsnitt 2.1. användes för variabler avseende hemma- och bortalag i samtliga försök. De uttryck jag testade för λ (se Avsnitt 2.3) var

$$\begin{aligned} \lambda_1 &= \frac{t}{k} \cdot 1\{t \leq k\} + 1\{t > k\} \\ \lambda_2 &= 1 - e^{-at}, \end{aligned} \tag{2}$$

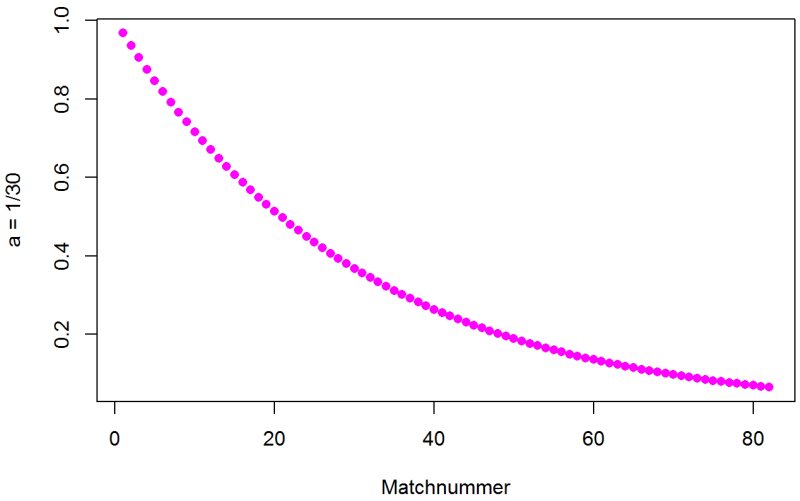
där t är antalet spelade matcher under säsongen och k och a konstanter.

Jag började med en regressionsmodell med samtliga variabler och testade att variera λ på WP.Tm och WP.Opp. Först testade jag med λ_1 för olika värden på k och sedan med λ_2 för olika värden på a. Sen gjorde jag samma sak med övriga variabler. Den slutliga modellen bestod av de förklarande variabler i Tabell 8 med tillhörande värde för λ.

Variabel	$\boldsymbol{\lambda}$	**Beskrivning**
DiffHth		Genomsnittlig målskillnad mot samma motståndare
sqrt(WinP.Tm)	$\lambda_2, a = 1/30$	Vinst% för hemmalaget
exp(GA.Tm)	$\lambda_2, k = 2$	Insläppta mål, hemmalag (i genomsnitt)
exp(GF.Opp)	$\lambda_1, k = 2$	Gjorda mål, bortalag (i genomsnitt)
GA.Opp	$\lambda_1, k = 3$	Insläppta mål, bortalag (i genomsnitt)
sqrt(GPP.Tm)		Antal spelade matcher i slutspelen förra säsongen, hemmalag
sqrt(SD)	$\lambda_1, k = 5$	Genomsnittlig skott% per match, hemmalag-bortalag
Pims.Opp		Genomsnittligt antal utvisningsminuter, bortalag

Tabell 8: Förklarande variabler i slutliga modellen. λ_1 och λ_2 är definierade enligt (2).

För de variabler med λ_1, innebär viktningen att föregående säsongs prestationer påverkar upp till match $k - 1$ under aktuella säsongen. För variabeln GA.Tm, med $\lambda_2 = 1 - e^{-2t}$, är den andelen av variabeln som utgörs av data från föregående säsong knappt 2%. I Figur 3 visas andelen av variabeln som utgörs av data från föregående säsong, d.v.s. $1 - \lambda_2$, då $a = 1/30$.



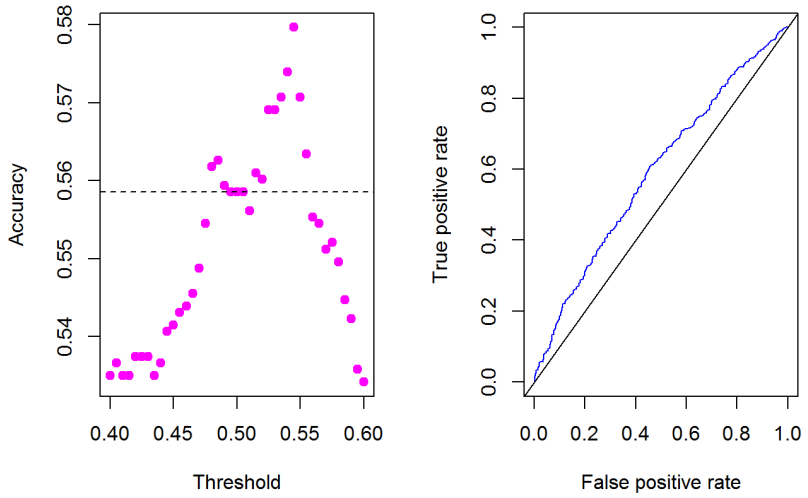
Figur 3: Andelen av variabeln sqrt(WinP.Tm) som utgörs av vinstprocenten föregående säsong, d.v.s.$1 - \lambda_2 = e^{-t/30}$

I Tabell 9 ser vi att klassifikationsförmågan är hyfsat låg för valideringsmängden samt att log loss inte är mycket lägre än i exemplet i Avsnitt 3.3. Precis som i Modell 1, har Modell 2 problem med att en stor andel förluster (71.7%) felaktigen klassificeras som vinster. Här klassificerades 76.3% av matcherna som vinster, medan den faktiska andelen bara var 52.9%. Av vinsterna klassificerades 19.7% felaktigen som förluster, vilket är nästan identiskt med Modell 1.

Mängd	**Klassifikationsförmåga**	**Log Loss**	**AUC**
Hela säsongen	55.9 %	0.68	0.591
Ordinarie- och Övertid	55.9%	0.68	0.594
Ordinarie speltid	57.9%	0.672	0.615

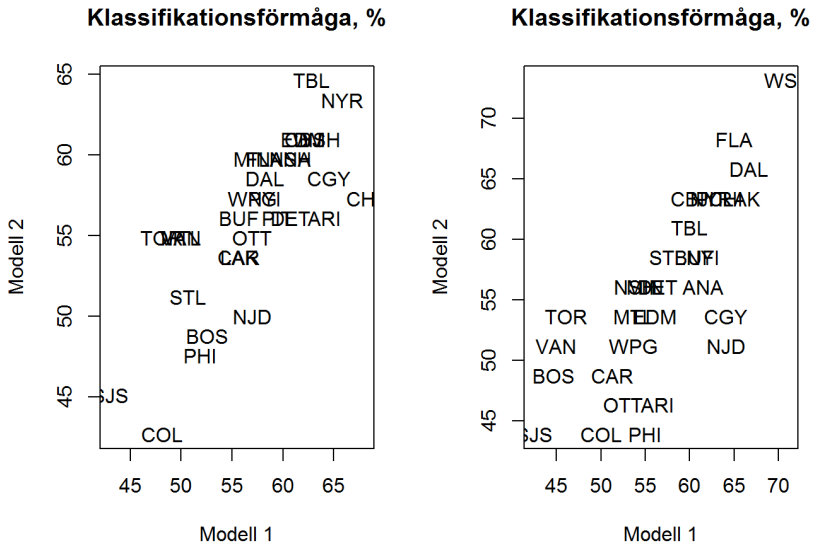
Tabell 9: Modell 2: Resultat på valideringsmängden.

I Figur 4 ser vi att brytpunkten 0.5 inte är den som ger högst andel korrekta klassifikationer. Då modellen har anpassats för att minimera log loss och inte maximera klassifikationsförmågan och det värde för brytpunkten som ger bäst klassifikationsförmåga inte skiljer sig avsevärt från 0.5, väljer jag att inte ändra på brytpunkten.



Figur 4: Andelen korrekta prediktioner för olika värden för brytpunkten (till vänster) och ROC-kurvan till höger.

I Figur 5 ser vi andelen korrekta prediktioner för Modell 1 och Modell 2 för varje lag, för samtliga matcher och för hemmamatcher. Båda modellerna ser ut att prediktera utfallen ungefär lika bra för lagen var för sig.



Figur 5: Klassifikationsförmåga för de olika lagen enligt Modell 1 och Modell 2, för samtliga undersökta matcher (till vänster) och för hemmamatcher (till höger).

Precis som för modell 1 undersökte jag hur koefficientskattningarna för modellen förändrades av att variera säsongerna i träningsmängden. Värden för klassifikationsförmåga, log loss och AUC finns i Tabell 10. Då den första säsongen, 2007/2008, exkluderades ökade andelen korrekta klassifikationer till 56.5%. Lägst värde för log loss (0.679) erhölls i alla försök förutom då träningsmängden endast bestod av säsongerna 2013/2014-2014/2015 samt 2014/2015.

Valideringsmängd	**Klassifikationsförmåga**	**Log Loss**	**AUC**
2007/2008-2014/2015	55.69 %	0.6792	0.5926
2008/2009-2014/2015	56.50 %	0.6792	0.5926
2009/2010-2014/2015	56.10%	0.6791	0.5932
2010/2011-2014/2015	55.69 %	0.6789	0.5924
2011/2012-2014/2015	55.61%	0.6793	0.5911
2012/2013-2014/2015	55.04%	0.6794	0.5919
2013/2014-2014/2015	55.69%	0.6797	0.5876
2014/2015	56.18%	0.6798	0.5875

Tabell 10: Modell 2: Resultat på validerings- och testmängd då träningsmängden reduceras.

I Tabell 11 och Tabell 12 finns koefficientskattningarna, standardfelelen, z-värdet och p-värdet för Modell 2, när träningsmängden består av säsongerna 2006/2007-2014/2015 samt 2008/2009-2014/2015.

Variabel	$\hat{\boldsymbol{\beta}}$	**Standardfel**	**z-värde**	**p-värde**
Intercept	-9.279949	0.944603	-9.824	$< 2e-16$
DiffHth	0.018362	0.016256	1.130	0.258669
sqrt(WinP.Tm)	1.879059	0.401014	4.686	$2.79e-06$
exp(GA.Tm)	-0.000856	0.001316	-0.650	0.515372
exp(GF.Opp)	-0.008230	0.002374	-3.466	0.000528
GA.Opp	0.153749	0.043065	3.570	0.000357
sqrt(GPP.Tm)	0.021658	0.014171	1.528	0.126427
sqrt(SD)	11.103706	1.369377	8.109	$5.12e-16$
Pims.Opp	-0.001320	0.001847	-0.714	0.474961

Tabell 11: Förklarande variabler i slutliga modellen, skattade koefficienter, standardfel, z-värde och p-värde.

Variabel	$\hat{\boldsymbol{\beta}}$	**Standardfel**	**z-värde**	**p-värde**
Intercept	$-10.50e$	$1.096e+00$	-9.575	$< 2e-16$
DiffHth	0.008179	0.01812	0.451	0.651636
sqrt(WinP.Tm)	1.704	0.4509	3.779	0.000157
exp(GA.Tm)	-0.0007752	0.001439	-0.539	0.590203
exp(GF.Opp)	-0.008034	0.002931	-2.742	0.006116
GA.Opp	0.1592	0.05072	3.139	0.001696
sqrt(PlayoffGP.Tm)	0.02714	$1.598e-02$	1.699	0.089340
sqrt(SD)	12.96	1.5990	8.106	$5.23e-16$
Pims.Opp	-0.002413	0.002567	-0.940	0.347294

Tabell 12: Koefficientskattningar för förklarande variabler i modellen med reducerad träningsmängd (2008/2009-2014/2015)

5 Resultat

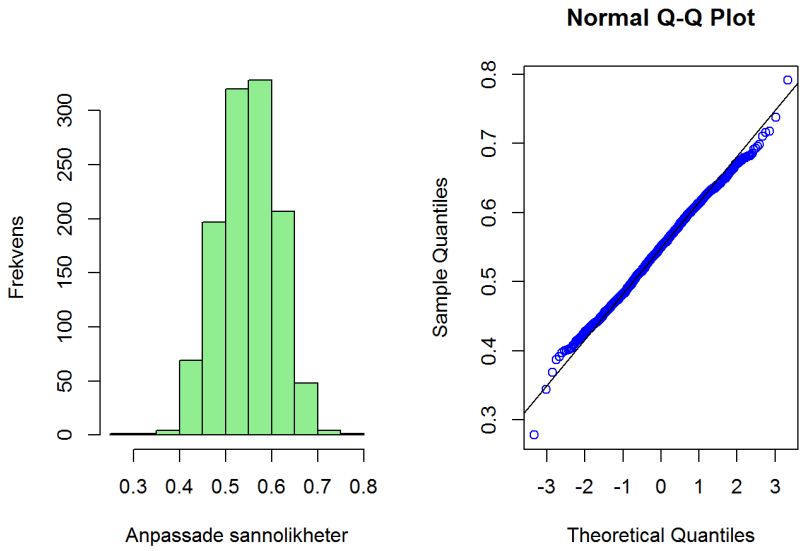
5.1 Modell 1

Den slutliga modellen lyckades prediktera rätt 56.8% av matcherna under säsong 2016/2017. I Tabell 13 ges värden för klassifikationsförmåga, log loss och AUC för olika delmängder av säsogen.

Mängd	**Klassifikationsförmåga**	**Log Loss**	**AUC**
Hela säsongen	56.8 %	0.677	0.588
Ordinarie- och Övertid	57.6%	0.674	0.599
Ordinarie matchtid	58.8%	0.672	0.6

Tabell 13: Modell 1: Resultat på testmängden.

Andelen förluster som felaktigen klassificerades som vinster var 72.4% och andelen vinster som felaktigen klassificerats som förluster var 19.3%. 76.9% av matcherna klassificerades som vinster, medan bara 54.9% faktiskt slutade med att hemmalaget vann.



Figur 6: Histogram och QQ-plot över anpassade sannolikheter.

I histogrammet samt i QQ-plotten i Figur 6 ser vi att sannolikheterna ser normalfördelade. Medelvärdet är 0.549 och standardavvikelsen 0.063. I Tabell 14 beskrivs andelen felaktiga klassificeringar för modellen då koefficienterna från den oreducerade samt den reducerade träningsmängderna användes.

Klassifikation	**Oreducerad testmängd**	**Reducerad testmängd**
Korrekt	56.8 %	57.3%
Felaktiga vinster	72.4 %	67.1%
Felaktiga förluster	19.3%	22.7%
Andel vinster	76.9%	72.7%

Tabell 14: Andel korrekta klassifikationer, andel av förlusterna som felaktigen klassificerats som vinster, andel av vinsterna som felaktigen klassificerats som förluster och andel av matcherna som klassificerades som vinster.

Förutom att andelen av förluster som felaktigen klassificerats som vinster minskade med 5.3 procentenheter, ökande andelen felaktiga klassificeringar av vinster med 3.4 procentenheter. Andelen korrekta klassifikationer ökade från 56.8% til 57.3%.

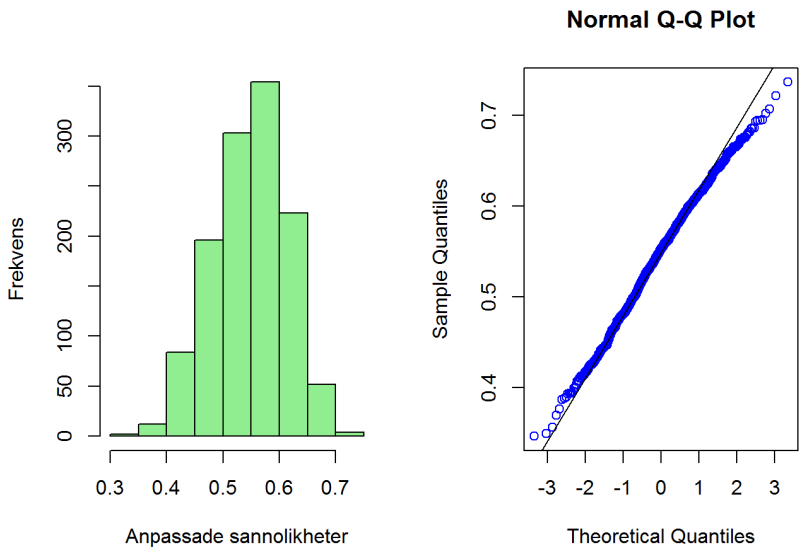
5.2 Modell 2

Andelen korrekta klassifikationer för resultaten i testmängden var ganska låg för båda modellerna, men något högre för Modell 2, som fick 58.5% av prediktionerna rätt. I Tabell 15 redovisas värdena för klassifikationsförmåga, log loss och AUC för olika delmängder av testmängden.

Mängd	Klassifikationsförmåga	Log Loss	AUC
Hela säsongen	58.5 %	0.673	0.592
Ordinarie- och Övertid	59.3%	0.669	0.608
Ordinarie speltid	60.6%	0.666	0.615

Tabell 15: Modell 2: Resultat på testmängden.

I Figur 7 ser vi att de anpassade sannolikheterna ser normalfördelade ut. Medelvärde och standardavvikelse för sannolikheterna är 0.548 respektive 0.064. Fördelningen för de anpassade sannolikheterna i båda modellerna är alltså väldigt lika

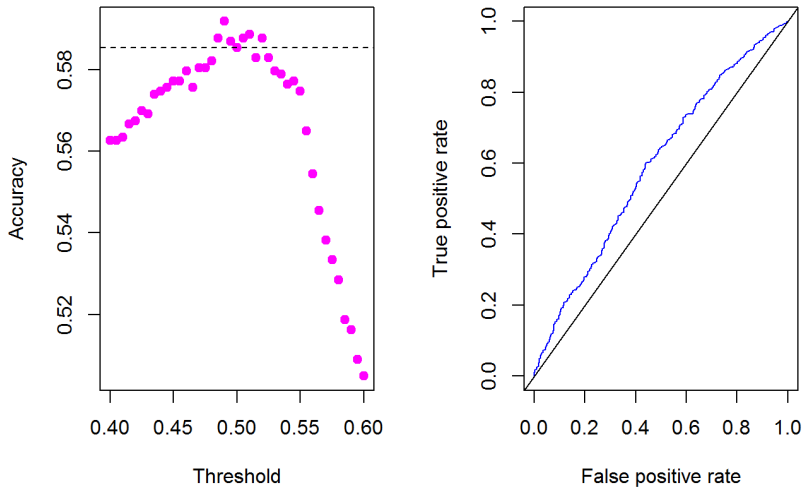


Figur 7: Anpassade sannolikheter.

Modellen med den reducerade träningsmängden gav en högre prediktionsförmåga, 59.2%. I Tabell 16 har vi andelen förluster som felaktigen klassificerats som vinster, andelen vinster som felaktigen klassificerats som förluster samt andel matcher som klassificerats som vinster för modellen med full träningsmängd och modellen med reducerad träningsmängd. Andelen vinster för hemmalaget under säsongen 2016/2017 var 55.9%.

Klassifikation	Oreducerad testmängd	Reducerad testmängd
Korrekt	58.5 %	59.2%
Felaktiga vinster	69.9%	68.8%
Felaktiga förluster	19%	18.8%
Andel vinster	76.1%	75.8%

Tabell 16: Andel korrekta klassifikationer, andel av förlusterna som felaktigen klassificerats som vinster, andel av vinsterna som felaktigen klassificerats som förluster och andel av matcherna som klassificerades som vinster.



Figur 8: Klassifikationsförmåga för olika värden på brytpunkten till vänster och ROC-kurva för testmängden till höger.

Som vi ser till vänster i Figur 8, hade ett högre värde på brytpunkten inte gett en mycket högre andel korrekta klassifikationer.

6 Diskussion och slutsats

Båda modellerna har predikterat många fler vinster än förluster för både validering- och testmängdera. Hemmalaget vinner visserligen fler matcher än de förlorar i genomsnitt, men vinstfrekvensen överskattas kraftigt av båda modellerna.

När de koefficientskattningar som modellerna med de reducerade träningsmängderna användes ökade prediktionsförmågan något, från 56.8% till 57.3% för Modell 1 samt från 58.5% till 59.2% för Modell 2. För båda modellerna minskade andelen förluster som felaktigen klassificerats som vinster och för Modell 2 minskade även andelen vinster som felaktigen klassificerades som förluster. För Modell 1

ökade däremot andelen av vinsterna som felaktigen klassificerades som förluster när reducerade träningsmängden användes.

Av de två anpassade modellerna har alltså Modell 2 lyckats bäst med att prediktera matchresultaten för säsong 2016/2017. Prediktionsförmågan var dock lägre än de som bl.a. erhölls av Weissbock och Pischedda, men något bättre än slumpen eller om man skulle gissat att hemmalaget alltid vann. Vidare utveckling av Modell 2 eller andra viktade modeller skulle därför vara intressant. Exempelvis skulle fler transformationer av variablerna kunna undersökas.

Eftersom både log loss och klassifikationsförmågan blev bättre då vissa av de tidiga säsongerna uteslöts, hade det varit intressant att anpassa en modell på de reducerade träningsmängderna om det funnits tid.

Ett alternativ till enkel logistisk regression hade varit att dela upp matchutfallet i 6 kategorier beroende på om hemmalaget vann eller förlorade i ordinarie matchtid, övertid eller straffar och använda multinomiell logistisk regression. Eller så hade jag kunnat använda målskillnaderna, möjligen uppdelade i olika intervall, som responsvariabel. Alternativt hade linjär diskriminantanalys kunnat användas istället för logistisk regression.

Referenser

[1] Weissbock, J. Forecasting Success in the National Hockey League using In-Game Statistics and Textual Data

[2] Thomas, A.C. (2007) Inter-arrival Times of Goals in Ice Hockey. *Journal of Quantitative Analysis in Sports 3.*

[3] Thomas, A. C., Ventura, S. L. , Jensen, S. T. & Ma, S. (2013) Competing process hazard function models for player ratings in ice hockey. *The Annals of Applied Statistics 3, s.1497-1524.*

[4] Routley, Kurt Douglas. (2015) A Markov Game Model for Valuing Player Actions in Ice Hockey

[5] Gramacy, R. B., Jensen, S.T. & Taddy, M. (2013) Estimating player contribution in hockey with regularized logistic regression. *Journal of Quantitative Analysis in Sports 1.*

[6] Weissbock, J., Viktor, H. & Inkpen, D. Use of Performance Metrics to Forecast Success in the National Hockey League

[7] Pischedda, Gianni (2014) Predicting NHL Match Outcomes with ML Models. *International Journal of Computer Applications 9. s.15-22.*

[8] Hockey Reference. URL www.hockey-reference.com

[9] Schuckers,M & Macdonald, B. (2014) Accounting for Rink Effects in the National Hockey League's Real Time Scoring System

[10] Beaudoin,D., Schulte, O. & Swartz, T.B. (2016) Biased Penalty Calls in the National Hockey League. *Statistical Analysis and Data Mining*

[11] Gill, Jeff. (2001) Generalized linear models.*SAGE Publications, Inc..*

[12] Myers, R. H., Montgomery, D. C., Vining, G. G. & Robinson, T.J. (2012) Generalized Linear Models: with Applications in Engineering and Sciences, s.4-10,2nd ed. *Hoboken : John Wiley & Sons.*

[13] Kaggle, www.kaggle.com/wiki/LogLoss

[14] Lobo, J.M., Jiménez-Valverde, A.& Real, R. (2007) AUC: a misleading measure of the performance of predictive distribution models. *Global Ecology and Biogeography.*

[15] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL www.R-project.org