Prediktera hörnor i första halvleken



Isaac Högfeldt

EC Utbildning

Projekt i data science

202510

Abstract

This report outlines the development of predictive models for estimating first-half corners for home and away football teams using match-related betting data. The dataset was split into training, validation, and test sets to ensure robust evaluation. Multiple regression models—including Linear Regression, Random Forest, and Gradient Boosting—were trained and assessed using Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). Feature importance analysis guided model selection, with Gradient Boosting chosen for predicting away team corners due to its superior handling of key features. The final models were validated on unseen data, visualized through prediction distributions, and saved for future deployment. The results offer insights into the relationship between betting odds, line totals, and corner outcomes, demonstrating the predictive strength of the selected models.

Innehållsförteckning

[Abstract 2](#_heading=h.gjdgxs)

[Innehållsförteckning 3](#_heading=h.ttyc5nvqhgq4)

[1 Inledning 4](#_heading=h.3znysh7)

[1.1 Syfte och Mål 4](#_heading=h.2et92p0)

[2 Teori 5](#_heading=h.tyjcwt)

[2.1 Machine learning 5](#_heading=h.f4h18ax2ziz0)

[3 Metod 7](#_heading=h.3rdcrjn)

[4 Resultat och Diskussion 8](#_heading=h.26in1rg)

[5 Slutsatser 10](#_heading=h.lnxbz9)

6 [Självutvärdering 11](#_heading=h.1ksv4uv)

[Appendix A 12](#_heading=h.44sinio)

[Källförteckning 13](#_heading=h.2jxsxqh)

1. Inledning

Att förutsäga utfall i sportevenemang är av stort intresse för både analytiker, spelbolag och sportentusiaster. I denna studie fokuseras på att utveckla prediktiva modeller för att uppskatta antalet hörnor under första halvlek för både hemmalag och bortalag i fotbollsmatcher. Genom att använda bettingrelaterade data kan modellerna ge värdefulla insikter om matchdynamik och potentiella utfall. Bettingdata erbjuder en rik källa till information, då den reflekterar marknadens kollektiva förväntningar och kan användas för att identifiera mönster och samband.

En särskilt intressant idé som nämns i Masaru Kanemotos bok *Winning Sports Betting* är att man kan genom att använda odds från ett spelbolag med hög volym och därmed hög marknadsinflytande, kan härleda så kallade "derivat" av ett spel vilket i denna rapportens fall är hörnor i en del av matchen.

Denna rapport presenterar utvecklingen av regressionsmodeller som tränats på historiska data, med syftet att förutsäga antalet hörnor för respektive lag. Modellerna utvärderas med hjälp av felmått som MAE och RMSE, och analys av feature importance används för att välja den mest lämpliga modellen för varje lag.

* 1. Syfte och Mål

Syftet är att utveckla prediktiva modeller för att uppskatta antalet hörnor under första halvlek i fotbollsmatcher genom att använda bettingrelaterad data som källa till insikter om matchdynamik och utfall. Ytterligare är syftet att undersöka hur marknadsbaserade odds kan användas för att härleda värdefulla spelderivat.

Mål:

* Hämta relevant historisk datan kring odds historik för de 5 stora fotbolls-ligorna.
* Skapa separata regressionsmodeller för hemmalag och bortalag baserat på historisk data.
* Utvärdera modellernas prestanda.
* Välja den mest lämpliga modellen för respektive lag.

1. Teori
   1. Machine learning

* **Supervised learning:** Är en maskininlärningsmetod där modellen tränas på en dataset som innehåller både indata (features) och tillhörande korrekta svar (labels). Målet är att modellen ska kunna lära sig sambandet mellan indata och labels för att sedan kunna göra korrekta förutsägelser på ny data. (Wikipedia (a), 2025)
* **Klassificeringsproblem och regrissionsproblem:** Inom maskininlärning delas problem ofta in i två huvudsakliga kategorier: klassificeringsproblem och regressionsproblem. Skillnaden mellan dessa typer av problem är avgörande för att välja rätt maskininlärningsmodell och utvärderingsmetodik.
* Klassificering: Klassificering innebär att förutsäga en diskret kategori eller klass. Exempel kan vara att klassificera en bild på en handskriven siffra utifrån vilken siffra det är.
* Regression: Regression handlar istället om att förutsäga ett kontinuerligt numeriskt värde. Exempel kan inkludera att förutsäga huspriser baserat på data som storlek, läge och byggnadsår.

(Wikipedia (a), 2025)

* **Tränings-, validerings- och test- dataset:** När en dataset används för maskininlärning delas den ofta upp i tre delar:
* Träningsset**:** Används för att träna modellen och justera dess parametrar.
* Valideringsset**:** Används för att utvärdera modellen och finjustera hyperparametrar utan att påverka träningsdata. Här finns valet att kombinera tränings- och valideringsdatan för en slutgiltig träning.
* Testset**:** Hålls separat och används enbart för att slutligt bedöma modellens prestanda, vilket ger en objektiv mätning av dess generaliseringsförmåga på ny data.

(Wikipedia (a), 2025)

* **Korrelationsmatris:** En matris som visualiserar korrelationen mellan alla variabler i ett dataset, vilket hjälper till att identifiera potentiell multikollinearitet. (Wikipedia (b), 2025)
* **Root Mean Squared Error (RMSE):** Mäter den genomsnittliga storleken på felen mellan de förutsagda och de faktiska värdena. Ett lägre RMSE indikerar bättre noggrannhet. (Wikipedia (c), 2025)
* **Feature importance:** Feature importance är ett mått på hur mycket varje indata-variabel bidrar till modellens prediktioner. Det används ofta inom träd-baserade modeller som Random Forest och Gradient Boosting för att identifiera vilka variabler som har störst påverkan på resultatet. (Wikipedia (d), 2025)
* **Overfitting:** Overfitting uppstår när en modell lär sig detaljer och brus i träningsdatan så väl att den presterar dåligt på ny, osedd data. (Wikipedia (e), 2025)
* **Efficient Market Hypothesis och spelmarknaden:** Inom sportsbetting används ofta antagandet att spelmarknaden är relativt effektiv, vilket innebär att odds från spelbolag reflekterar all tillgänglig information om en match. Detta bygger på den ekonomiska teorin Efficient Market Hypothesis, som hävdar att priser på en marknad inkorporerar all relevant information och därmed gör det svårt att konsekvent slå marknaden genom analys eller strategi. I bettingkontext innebär det att odds från spelbolag med högt marknadsinflytande kan betraktas som en form av kollektiv förväntan. (Wikipedia (f), 2025)

1. Metod

Studien genomfördes enligt följande steg:

1. **Datainläsning och förbehandling:**

* Insamling av matchdata från [corner-stats.com](http://corner-stats.com) med fokus på bettingrelaterade variabler såsom odds och linjer för målspel och asiatiska handikapp.
* Val av relevanta features för modellträning: close\_line\_total, close\_odds\_total\_over, close\_odds\_total\_under, close\_odds\_ah\_1, close\_odds\_ah\_2, close\_line\_ah\_1, close\_line\_ah\_2.
* Ingen ytterligare transformation eller encoding krävdes då samtliga variabler var numeriska och färdiga för modellering.

1. **Uppdelning av data:**

* 70 % (1652 rader) av datan användes för träning, resterande 30 % delades upp i validerings- och testuppsättningar (354 rader vardera).
* Separata målvärden användes för hemmalag och bortalag, vilket möjliggjorde träning av två parallella modeller.

1. **Modellträning:**

* Tre regressionsmodeller testades: Linear Regression, Random Forest Regressor och Gradient Boosting Regressor.
* Modellerna tränades på träningsdatan och utvärderades på valideringsdatan.
* Feature importance analyserades för att utforska om intuitivt viktiga variabler påverkade prediktionen av hörnantal. Intuitivt bör linor bör vara väldigt likgiltigt odds då linan visar antalet som oddsen speglar.

1. **Modellval:**

* Gradient Boosting valdes för bortalagets modell på grund av bättre hantering av icke-linjära samband.
* Linear Regression valdes för hemmalagets modell baserat på stabil prestanda och tolkningsbarhet.

1. **Utvärdering:**

* Modellernas prestanda mättes med hjälp av MAE (Mean Absolute Error) och RMSE (Root Mean Squared Error).
* Korrelationsmatriser användes för att undersöka sambandet mellan features och målvärden.
* Visualiseringar av predikterade kontra faktiska hörnantal skapades för både hemmalag och bortalag.
* Histogram användes för att analysera fördelningen av prediktioner och identifiera eventuella extremvärden.

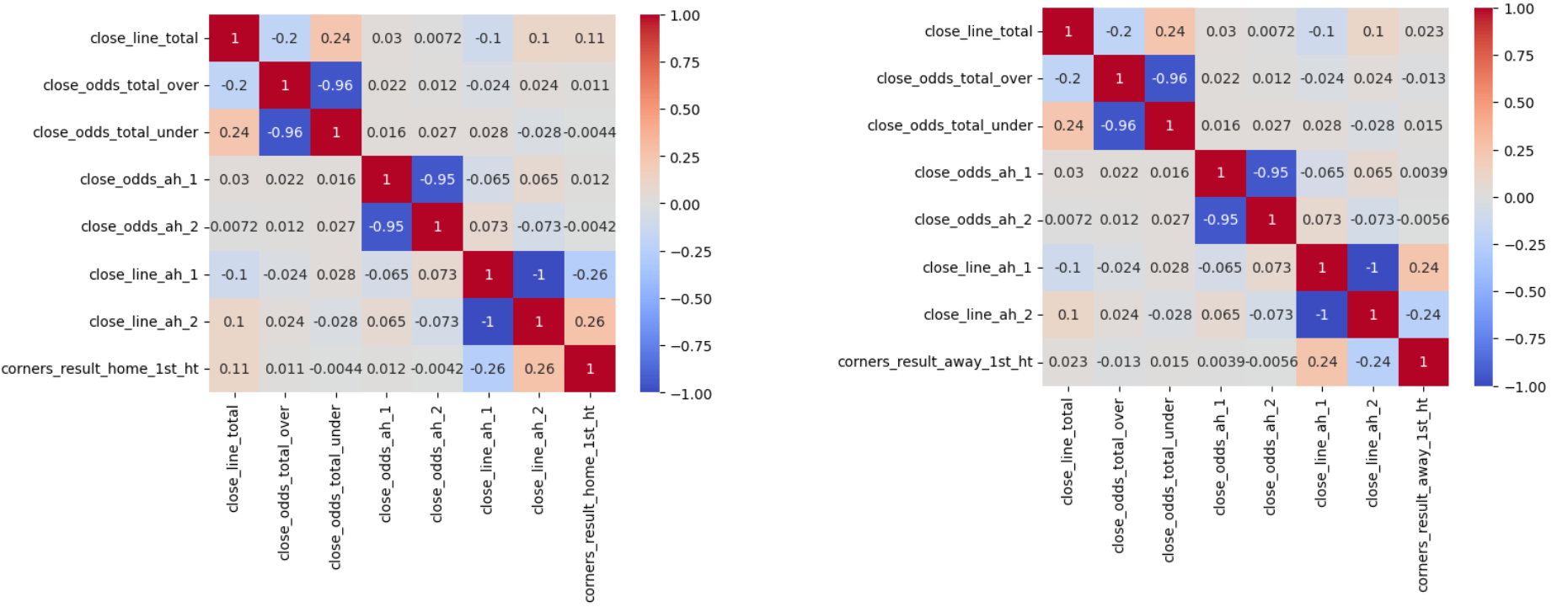
1. **Export:**

* De bästa modellerna sparades som .joblib-filer för framtida användning.

1. Resultat och Diskussion

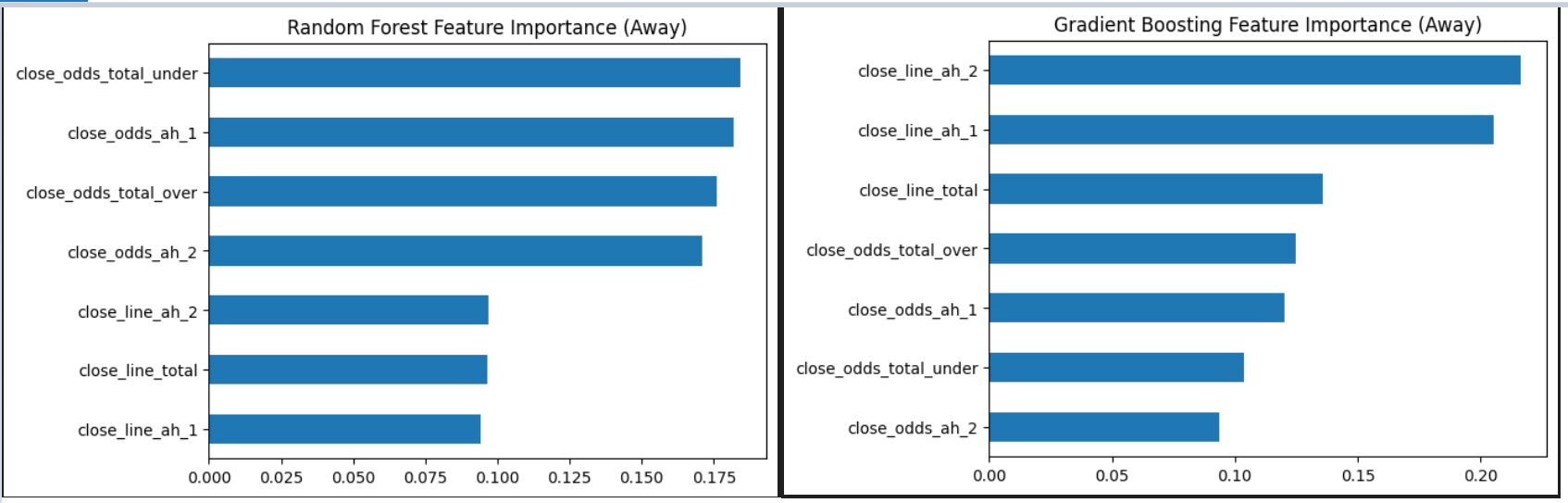
Modellerna utvärderades med hjälp av måtten Mean Absolute Error (MAE) och Root Mean Squared Error (RMSE). För hemmalaget visade den slutgiltiga modellen (Linear Regression) en **RMSE på 1.73** och en **MAE på 1.38**, medan bortalagets modell (Gradient Boosting) presterade med en **RMSE på 1.61** och en **MAE på 1.27**. Dessa resultat indikerar att modellerna har en förmåga att förutsäga hörnantalet under första halvlek, med relativt låg genomsnittlig avvikelse från de faktiska värdena.

För att ytterligare utvärdera modellernas prestanda genomfördes ett paired t-test där varje modells felvärden jämfördes med en baslinjemodell som enbart förutsäger medelvärdet. Resultatet visade att hemmalagets modell var statistiskt signifikant bättre än baslinjen med ett **p-värde på 0.0390**, vilket understiger signifikansnivån på 0.05. Detta stärker slutsatsen att modellen har verklig förklaringskraft och inte enbart följer slumpmässiga mönster. För bortalagets modell fick ett **p-värde på 0.2107**, vilket inte är statistiskt signifikant. Trots detta presterade modellen bättre än baslinjen sett till både MAE och RMSE. Jag är osäker hur detta resultat skall tolkas, vid närmare analys av “predictions vs actual” syns en skillnad mellan hemma och borta modellen, nämligen att borta modellen gissar väldigt centrerat runt 2 medan hemma modellen har mer normalfördelade gissningar. Detta kan vara ett tecken att borta modellen är “over-fitted”. Kanske var det ett misstag att välja bort linear regression för borta modellen.



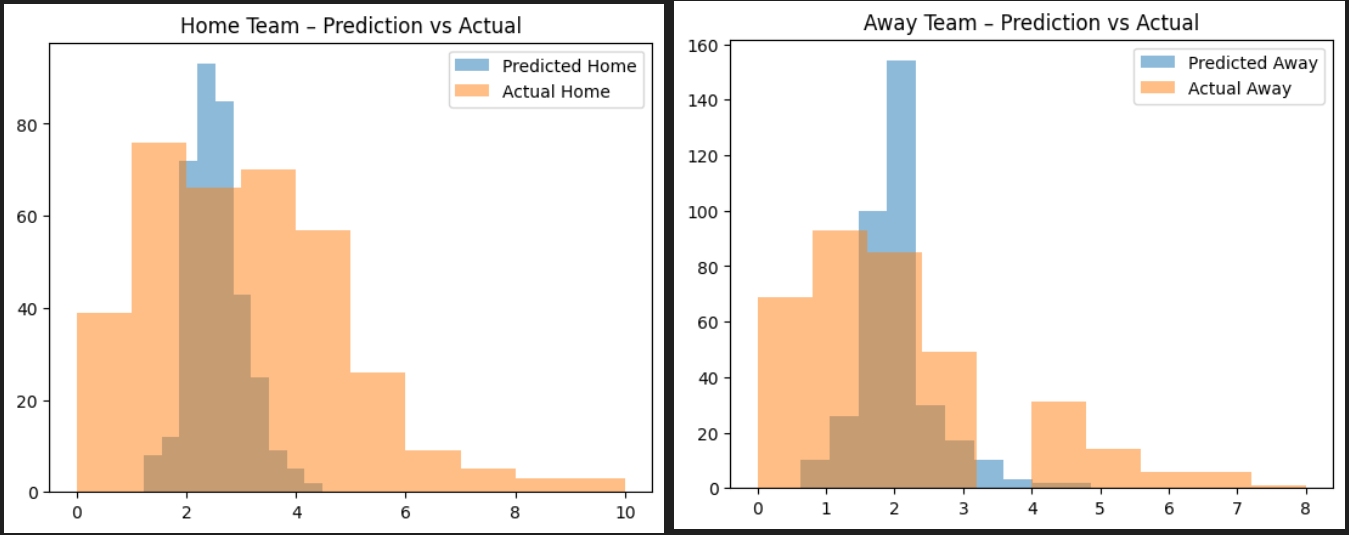
***Figur 1,*** *korrelationsmatriser för de två mål-variablerna*

Korellationsmatriserna visar sambandet mellan de olika features och hörnantalet för respektive lag. Det kan observeras att vissa variabler, såsom close\_odds\_total\_under och close\_line\_ah\_1, har tydliga samband med hörnantalet för hemmalaget, medan close\_line\_total visar svagare korrelation med bortalagets hörnor. Detta bekräftar att vissa features har olika påverkan beroende på lagtyp.

**

***Figur 2,*** *Feature importance graferna för Random Forest och Gradient Boosting*

Feature importance graferna visar att variabler som close\_odds\_ah\_2 och close\_line\_ah\_2 har stor betydelse för prediktionerna, särskilt för bortalaget. Denna insikt användes för att välja Gradient Boosting som slutgiltig modell för bortalaget, då den bättre hanterar komplexa och icke-linjära samband.



***Figur 3,*** *Prediction vs actual corners graferna för de slutgiltiga två valda modellerna*

Grafen över predikterade kontra faktiska hörnantal visar att det inte finns några totalt felaktiga gissningar såsom negativa eller överdrivet låga kontra höga gissningar.

Sammantaget visar resultaten att bettingdata kan användas effektivt för att förutsäga hörnantal i fotbollsmatcher. Genom att analysera korrelationer, feature importance och prediktionsfördelningar kan man identifiera vilka variabler som är mest relevanta, och därmed bygga modeller. Däremot bör mer analysering och förhoppningsvis förbättringar tillämpas för att förbättra modellernas prestanda.

1. Slutsatser

Den utvecklade regressionsmodellen har visat sig vara kapabel att förutsäga antalet hörnor under första halvlek för både hemma- och bortalag, baserat på bettingrelaterad data. Modellernas prestanda, mätt med MAE och RMSE, visar att de genererar användbara uppskattningar, även om viss variation mellan predikterade och faktiska värden kvarstår. Statistisk analys visade att hemmalagets modell presterade signifikant bättre än en baslinjemodell, medan bortalagets modell visade tecken på överanpassning, vilket bör beaktas i framtida förbättringar.

Korrelationsanalysen och feature importance-visualiseringarna identifierade flera variabler med stark påverkan på hörnantalet, särskilt inom asiatiska handikapp och totalsummor. Dessa insikter stämmer överens med intuitiva antaganden om hur marknadsdata reflekterar matchdynamik.

För att vara extra tydlig kring modellernas begränsningar: eftersom hörnantalet i en enskild match påverkas starkt av slumpmässiga faktorer, kommer ingen modell kunna förutsäga utfallen med full träffsäkerhet. Ett mer realistiskt mål för fortsatt utveckling vore att använda de predikterade hörnantalet för att beräkna sannolikheter för olika linor—till exempel över 2.5 hörnor för ett lag i första halvleken—och jämföra dessa sannolikheter med spelmarknadens odds. Givet att spelmarknaden är relativt effektiv tack vare sitt höga marknadsinflytande, skulle en modell som genererar sannolikheter nära genomsnittet av flera spelbolags odds kunna anses vara välkalibrerad och användbar för praktisk tillämpning.

1. Självutvärdering
2. Utmaningar du haft under arbetet samt hur du hanterat dem.

Min största utmaning var att utifrån datan jag hade få ut någon meningsfull modell / output. Jag började med odds data från 1x2, över under mål samt över under gula/röda kort som ses i “old” mappen på github. Min tanke var att genom att träna en modell på denna data kunna skapa en modell som kommenterade matchernas förutsätta matchbild. Jag insåg dock mitt misstag att det är väldigt svårt att nå det målet genom unsupervised learning. Hur jag än vred och vände hamnade jag bara med en plan som lika gärna hade kunnat använda regel-baserade kommentarer istället för en maskininlärd modell.   
  
Ytterligare en utmaning är väl att jag inte riktigt fullbordade modellen då jag vet att det finns rum för förbättringar utifrån min analys. Jag hade gärna dykt djupare i analysen men jag vet inte riktigt vad nästa steg skulle vara tyvärr.

1. Vilket betyg du anser att du skall ha och varför.

G, jag tycker att jag har fullbordat uppgiftskraven men för VG tror jag att det hade behövts en mer kunskaps kring analysering av modellval samt slutliga test-resultaten.

1. Något du vill lyfta fram till Antonio?

Väldigt roligt sista projekt, jag känner verkligen att jag lärt mig mycket kring maskininlärning med python. Ska jag vara ärlig så inser jag nu att jag hållit mig kvar i min python “comfort-zone” istället för att utforska / lära mig mer om Power BI, R eller SQL. Hade jag gjort om de senaste kurserna hade jag valt att fokusera mer på de verktygen för att inte bara förbättra maskininlärning och datahantering i python.

Appendix A

[Github](https://github.com/IsaacHog/ec-data-science-24/tree/main/projekt_i_data_science)

Källförteckning

1. [Corner-stats.com](http://corner-stats.com) (2023). *Database of fotball stats*. Hämtad april 21, 2021.
2. Wikipedia contributors (a) (2025). *Machine learning*. Hämtad oktober 19, 2025, från Wikipedias sida <https://en.wikipedia.org/wiki/Machine_learning>.
3. Wikipedia contributors (b) (2025). *Correlation*. Hämtad oktober 19, 2025, från Wikipedias sida <https://en.wikipedia.org/wiki/Correlation>.
4. Wikipedia contributors (c) (2025). *Root mean square deviation*. Hämtad oktober 19, 2025, från Wikipedias sida <https://en.wikipedia.org/wiki/Root_mean_square_deviation>.
5. Wikipedia contributors (d) (2025). *Feature selection*. Hämtad oktober 21, 2025, från Wikipedias sida <https://en.wikipedia.org/wiki/Feature_selection>
6. Wikipedia contributors (e) (2025). *Overfitting*. Hämtad oktober 21, 2025, från Wikipedias sida <https://en.wikipedia.org/wiki/Overfitting>
7. Wikipedia contributors (f) (2025). *Efficient-market hypothesis*. Hämtad oktober 21, 2025, från Wikipedias sida <https://en.wikipedia.org/wiki/Efficient-market_hypothesis>