Predicția echipei câștigătoare a unui meci de fotbal

- Raport final –

Autori: Caragea Matei-Ioan

Barbu Andrei-Cătălin

1. Introducere / Context

Fotbalul este unul dintre cele mai populare sporturi la nivel global, generând un volum imens de date statistice la fiecare meci. Predicția corectă a rezultatului unei partide (victoria gazdelor, egal sau victoria oaspeților) este o problemă complexă din cauza naturii stocastice a sportului și a numărului mare de variabile care pot influența scorul final.

În mod tradițional, predicțiile sunt realizate de experți umani sau case de pariuri, bazându-se pe intuiție și experiență. Acest proiect își propune să elimine subiectivismul uman prin dezvoltarea unui sistem automatizat care analizează modele statistice obiective pentru a determina cea mai probabilă ieșire a unui eveniment sportiv.

Obiectivul principal este construirea unui model de clasificare (Machine Learning) capabil să prezică câștigătoarea unui meci de fotbal. Proiectul abordează întregul flux de lucru specific științei datelor:

1. **Colectarea datelor (Data Mining):** Extragerea datelor brute din surse web.
2. **Analiza Exploratoare a Datelor (EDA):** Investigarea inițială a setului de date pentru a descoperi tipare, a identifica anomalii și a vizualiza relațiile dintre variabile.
3. **Procesarea datelor:** Curățarea și transformarea statisticilor în feature-uri relevante.
4. **Modelarea predictivă:** Antrenarea algoritmilor pe date istorice.

2. Setul de Date

În urma procesului de scraping și validare inițială, a rezultat un set de date final compus din 1.596 de înregistrări. Fiecare intrare reprezintă un meci unic de fotbal disputat în campionatul intern, conținând atât rezultatul final, cât și vectorul complet de statistici tehnice asociate (posesie, șuturi, etc.). Acest volum de date este suficient pentru a reduce riscul de overfitting specific seturilor de date mici.

3. Analiza Exploratoare a Setului de Date

Această etapă a avut ca scop înțelegerea profundă a distribuției datelor și validarea ipotezelor statistice înainte de antrenarea modelului. Am utilizat un set de vizualizări grafice pentru a identifica dezechilibrele de clase, consistența setului de date și puterea predictivă a diverselor variabile.

Analiza distribuției rezultatelor pe întregul set de date a evidențiat un "Home Field Advantage" pronunțat. Procentajul victoriilor gazdelor (42.1%) este superior egalurilor (27.8%) și victoriilor oaspeților (30.0%). Acest dezechilibru natural al claselor (Class Imbalance) impune utilizarea unor metrici de evaluare care să nu fie biasate de clasa majoritară (precum F1-Score sau Precision/Recall), o acuratețe de tip "baseline" fiind deja de 42%.

S-a observat o variație semnificativă a numărului de înregistrări per echipă. În timp ce echipele de top (ex: CFR Cluj, FCSB) au peste 200 de meciuri, echipele nou-promovate au sub 5 înregistrări. Pentru a evita zgomotul statistic generat de eșantioanele mici, mediile mobile vor fi calculate doar acolo unde există un istoric relevant, sau se va aplica o regularizare a datelor pentru echipele cu puține meciuri.

Compararea mediilor statistice între echipele câștigătoare și cele învinse a permis separarea indicatorilor performanți de cei irelevanți:

1. **Predictori Puternici** (Șuturile pe Poartă): Există o diferențiere clară între câștigători (media 5.16) și învinși (media 3.12). Aceasta indică o corelație pozitivă puternică cu rezultatul final.
2. **Predictori Slabi** (Posesia și Faulturile): Analiza a infirmat mitul posesiei ca factor decisiv. Posesia medie este aproape identică între câștigători (49.8%) și învinși (50.2%), indicând o lipsă de cauzalitate directă (posesie sterilă). Similar, numărul de faulturi nu variază semnificativ (13.15 vs 12.73).

4. Preprocesarea Datelor

Transformarea datelor brute în atribute predictive s-a realizat cu biblioteca Pandas, urmărind strict cronologia evenimentelor pentru a evita data leakage.

Setul a fost ordonat cronologic, iar tipurile de date au fost standardizate. Valorile lipsă au fost tratate prin două metode specifice:

- Posesia: S-a aplicat o logică de completare (100% - valoarea adversarului) sau imputare cu 50% în lipsa ambelor valori.
- Statistici Tehnice: Restul valorilor lipsă au fost înlocuite cu mediana distribuției, metodă robustă la valorile extreme (*outliers*).

Pentru a prezice rezultatul, au fost create variabile care reflectă forma de moment și contextul competițional:

- Medii Mobile: S-a calculat media ultimelor 5 meciuri pentru fiecare indicator tehnic (șuturi, cornere, etc.). Astfel, modelul învață din forma recentă a echipei, nu din statistica meciului curent.
- Reconstrucția Clasamentului: Deoarece poziția în clasament lipsea din datele brute, a fost implementat un algoritm care a simulat desfășurarea sezonului, recalculând punctajul și golaverajul înaintea fiecărei etape pentru a determina locul exact ocupat de echipe

Variabila dependentă a fost discretizată în trei clase: 1 (Victorie Gazdă), 0 (Egal), 2 (Victorie Oaspete). Coloanele cu numărul de goluri ale meciului curent au fost eliminate din setul final de antrenare.

5. Modelarea Predictivă și Evaluarea Rezultatelor

Pentru faza de modelare, setul de date procesat a fost împărțit în set de antrenare (80%) și set de testare (20%). S-au antrenat și comparat trei algoritmi de clasificare, performanța acestora fiind evaluată prin intermediul Matricelor de Confuzie, care evidențiază capacitatea modelelor de a distinge corect între cele trei clase (1, 0, 2).

A. **Random Forest**: Acest model a demonstrat o tendință puternică de a favoriza clasa majoritară (Victoria Gazdelor). Deși are o rată impresionantă de predicție corectă a victoriilor gazdelor (78%), modelul eșuează aproape complet în detectarea rezultatelor de egalitate, clasificând corect doar 2.8% dintre acestea. Random Forest a "memorat" faptul că gazdele câștigă des și a maximizat acuratețea globală ignorând meciurile echilibrate.

B. **Gradient Boosting**: Spre deosebire de Random Forest, algoritmul Gradient Boosting a demonstrat cel mai mic bias, oferind o distribuție mult mai realistă a predicțiilor.

- Egaluri (X): A reușit să identifice corect 17% dintre remize, o performanță de 6 ori mai bună decât Random Forest.
- Oaspeți (2): A prezis corect 40% dintre victoriile oaspeților, fiind superior celorlalte modele la acest capitol.
- Gazde (1): Deși acuratețea pe gazde a scăzut la 64%, acest lucru indică un model care nu se bazează exclusiv pe avantajul terenului propriu, ci analizează activ feature-urile statistice.

C. **Regresie Logistică / SVM:** Al treilea model testat s-a situat la mijloc, având o rată de detecție a egalurilor de 11% și a victoriilor gazdelor de 70%. Performanța sa este inferioară Gradient Boosting în ceea ce privește sensibilitatea la clasele minoritare (0 și 2).

6. Concluzii și Selecția Modelului Final

Analiza matricelor de confuzie indică faptul că Gradient Boosting este cel mai robust algoritm pentru această problemă. În pariurile sportive, valoarea nu stă doar în prezicerea favoriților, ci în capacitatea de a identifica surprizele și meciurile strânse.

Deși Random Forest poate avea o acuratețe globală ușor mai mare, aceasta este "artificială", provenită din ignorarea remizelor. Gradient Boosting oferă cel mai bun echilibru între precizie și capacitatea de generalizare pe toate cele trei rezultate posibile.