Analiza valorii de piață a jucătorilor de fotbal

**Introducere**

Fotbalul este un sport vechi, apărut în urmă cu 158 de ani la Londra, unde s-a format prima asociație de fotbal, numită , în engleză, *Football Association.*

Ca oricare alt lucru, a evoluat pe parcursul anilor, devenind din ce în ce mai popular, extinzându-se și în afara granițelor britanice. Astăzi, fotbalul este un sport internațional, cu sute de echipe componente și zeci de mii de jucători activi.

Deoarece stârnește interesul multor oameni (Cupa Mondială din 2018 a avut un total de 3 milioane de spectatori în cele 64 de meciuri, iar finala a strâns 517 milioane de telespectatori, conform [1]; clubul Manchester United a vândut peste 3 milioane de tricouri în sezonul 2018/2019, generând venituri de peste 300 de milioane de euro doar din vânzarea lor, conform [2]), fotbalul nu mai este demult doar un sport, ci o afacere profitabilă, din care se pot genera profituri substanțiale (clubul FC Barcelona a generat venituri de 715 milioane de euro în sezonul 2019/2020, iar primele 10 echipe din top au generat un venit cumulat de peste 5 miliarde de euro, conform [3]) . Astfel, lumea fotbalului a atras atenția tot mai multor investitori care nu sunt neapărat fani ai fotbalului, dar au văzut în el o investiție profitabilă.

Un mijloc important de generare de venituri este vânzarea de jucători. Cluburile își doresc performanțe, iar pentru a reuși caută construirea echipei optime, în funcție de anumite criterii prestabilite. Acest lucru este greu de realizat doar cu proprii jucători, crescuți la academie, fiind nevoie de realizarea de transferuri de jucători de la alte echipe. Bineînțeles, acest proces implică sume din ce în ce mai consistente, în funcție de factori precum calitatea jucătorului, vârsta, poziția pe care joacă, potențialul lui etc.

Lucrarea de față își propune să răspundă la întrebarea *Care este prețul corect al unui jucător și de ce factori depinde el*?. Nimeni nu dorește să plătească prețuri *umflate*, iar ehipele nu-și doresc să renunțe la jucătorii de calitate pentru prețuri modice. Însă care este, de fapt, prețul corect? Cel propus de cumpărător sau cel cerut de vânzător? Bineînțeles, subiectivismul își pune amprenta puternic în astfel de situații. Astfel, crearea unui model care să elimine nota subiectivă și să păstreze criterii relevante, apropiate de adevăr, va ajuta la stabilirea prețurilor jucătorilor și la ameliorarea discuțiilor între părțile implicate. Cota de piață aproximată de model va putea fi folosită ca reper în negocieri.

Raportul propune 4 metode de analiză a prețurilor: modelul liniar (regresie), arbore de decizie, arbore de decizie construit cu bagging și arbore de decizie generat cu random forests. RMSE este indicatorul folosit pentru compararea acurateței predicțiilor, folosindu-se, de asemenea, grafice care să clarifice diferențele dintre modele.

**Setul de date**

Pentru analiza valorii prețurilor jucătorilor de fotbal profesioniști, s-a folosit un set de date extrase jocul FIFA 2020, un simulator al fotbalului, care cuprinde date reale și veridice despre jucători, fiind licențiate de FIFPro. Setul cuprinde informații despre 18278 de jucători, detaliate cu ajutorul a 104 variabile. Variabila țintă se numește *value\_eur,* adică valoarea jucătorilor în euro. Anumite atribute au fost eliminate, deoarece nu sunt relevante pentru realizarea analizei ori sunt date redundante. Procesul de curățare a implicat eliminarea a 53 de variabile, cum ar fi numele jucătorilor, numărul de pe tricou, poziția preferată la echipa națională etc. Poziția în teren influențează prețul jucătorilor, însă în setul de date coloana care conține informații despre acest aspect enumeră toate pozițiile preferate ale jucătorilor. Pentru a putea utiliza informația, a fost extrasă prima valoare enumerată pentru fiecare jucător, deoarece este cea principală, fiind salvată într-o variabilă nouă, numită *main\_position.* Variabilele *international\_reputation, skill\_moves (* nivelul abilității de a realiza diverse mișcări cu mingea, de la 1-5 *), weak\_foot (* nivelul abilității de a folosi piciorul secundar relativ la piciorul de bază, de la 1-5 *),* și  *main\_position* au fost factorizate, deoarece sunt date care clasifică jucătorii pe categorii. De asemenea, au fost eliminați portarii, deoarece au atribute diferite față de ceilalți jucători, astfel pot influența eronat analiza. Au fost păstrați spre analiză doar jucătorii de câmp.

Atributele rămase sunt: *overall*, care prezintă, pe o scară de la 1-99, calitatea jucătorului, fiind un atribut calculat în funcție de poziția în teren, ținând cont de atribute specifice fiecăreia, deci reprezintă un atribut obiectiv în compararea jucătorilor și de reputația internațională; de la *pace* la *goalkeeping\_reflexes* sunt atribute care detaliază calitatea jucătorilor, în funcție de diverși parametrii care, în funcție de poziție, sunt folosiți în determinarea *overall-*ului. Alte atribute sunt: *age, wage\_eur (*salariu săptămânal în euro*), international\_reputation, weak\_foot, skill\_moves, main\_position* și *potential (*calitatea potențială a jucătorilor pe care o pot atinge în viitor*).*

Pentru vizualizarea datelor rezultate s-a folosit librăria Data Explorer, care realizează un raport complet, prin prezentarea completitudinii observațiilor, histograme pentru fiecare atribute, harta de corelație, analiza componentelor principale și altele. Astfel, aproximarea valorii jucătorilor ia în calcul factori relevanți, obiectivi și folosește date complete pentru realizarea analizei.

**Rezultate și discuții**

**1. Modelul liniar**

În scopul explorării setului de date, am folosit librăria *Data Explorer*, pentru a genera un raport detaliat. Acest raport cuprinde informații vaste referitoare la setul de date, precum statistici de bază care conțin numărul de observații, de variabile, tipul lor, etc. De asemenea, mai sunt cuprinse în cadrul acestui raport histograme, *bar-charts,* QQ Plots ale variabilelor, precum și analize ale corelațiilor și a componentei principale.

Chart, line chart

Description automatically generated În urma analizei graficelor de tip QQ Plots, am observat o distribuție asemănătoare a variabilei *wage\_eur* cu cea a variabilei țintă, *value\_eur*. Această asemănare scoate în evidență o posibilă corelație între cele două variabile (poate fi observată în Figura 1). Alte variabile care ne așteptăm să influențeze valoarea jucătorilor sunt *age, potential* și *overall.* Bineînțeles, și alte variabile vor fi testate pentru a verifica existența unei corelații.

*Figura 1. QQ Plot*

Variabilele menționate anterior au fost testate singular în modele liniare, rezultând informații despre intensitatea legăturii cu variabila țintă. În urma testării, variabila *wage\_eur* explică 73,78% din varianța lui *value\_eur*, *overall* explică 41,47%, *potential* 33,81%, *age* 0,7%.

Pe baza acestui clasament, am construit modele adăugând variabilele în mod succesiv modelului liniar. De asemenea, au fost construite modele și pentru alte combinații de variabile posibile relevante, precum este exemplificat în figura următoare.

Application

Description automatically generated with medium confidence

*Figura 2*

*A picture containing text, receipt

Description automatically generated* În urma interpretării modelelor ilustrate anterior, am ajuns la concluzia că modelul WOA este cel mai performant deoarece variabilele prezente sunt toate relevante. Acest fapt este demonstrat prin valorile foarte mici (apropiate de 0) ale indicelui P-value, respingând posibilitatea unei legături întâmplătoare cu *value\_eur.* De asemenea, ipoteza nulă este respinsă deoarece valorile indicelui t-value sunt depărtate de 0. Acest model explică 79% din varianța lui *value\_eur* și are o eroare standard reziduală de 2.656.000. Observăm aceste aspecte în cadrul *Figurii 3.*

*Figura 3*

Performanța modelului WOA este confirmată și de coeficientul *AIC,* care evaluează calitatea acestuia pentru un anume set de date. Datorită valorii mici a coeficientului, modelul *WOA* are cel mai bun raport dintre numărul de variabile folosite și corectitudinea estimării valorilor variabilei țintă (vezi *Figura 4*).

A picture containing text, receipt

Description automatically generated

*Figura 4*

Eroarea de predicție (RMSE) a modelului WOA este de 2.383.658 euro.

Chart, histogram

Description automatically generatedPentru a evalua performanța modelului, am realizat două grafice, unul care prezintă atât valorile actuale, cât și cele previzionate ale variabilei dependente, iar celălalt distribuția față de 0 a diferenței dintre cele două valori. Valorile prezise sunt ilustrate cu culoarea albastră, iar cele actuale cu roșu.

*Figura 5*

Chart, scatter chart

Description automatically generated

*Figura 6*

Graficele din figurile 5, respectiv 6 ilustrează în mod intuitiv performanța modelului WOA, care poate fi apreciată ca fiind bună. În continuare, vom evalua alte modele, bazate pe arbori de decizie, pentru a prezenta comparativ performanța acestora.

**2. Arbore de decizie**

În continuare, am folosit ca metodă de evaluare a valorilor jucătorilor arborele de decizie. Pentru construirea lui, am folosit librăria ***rpart***, unde modelul a avut la dispoziție la fiecare pas tot setul de variabile pentru a determina variabila optimă în funcție de care să realizeze tăierea.

Modelul rezultat a realizat 7 tăieri, cu un cp (cost complexity) în valoare de 0,01. Eroarea de predicție (RMSE) a modelului *simpleTree* este de 1.740.406 euro.

Arborele rezultat este ilustrat în figura următoare. Observăm aspectul că variabila *overall* este cea mai importantă, deoarece este considerată variabila decisivă, fiind prezentă în primele 4 tăieri. Următoarea variabilă ca importanță a fost selectată cea de *potential*, fiind prezentă în ultimele 3 tăieri. Alegerea acestei variabile poate fi explicată prin faptul că potențialul unui jucător este întotdeauna mai mare sau egal cu calitatea jucătorului (*overall*). Prin urmare, un jucător foarte bun (cu o valoare a variabilei *overall* mare) va avea și o cotă de piață (*value\_eur*) mare, și un jucător tânăr cu un *overall* mai mic, dar cu un potențial ridicat are o cotă de piață considerabilă. În același timp, un jucător experimentat cu un *overall* mai mic, nu mai are potențial de creștere, deci va avea o cotă redusă, fapt prin care explicăm alegerea variabilei *overall* și nu a potențialului la cea de-a 4a tăiere.

Diagram

Description automatically generated

*Figura 7*

Pentru a evalua performanța modelului, am realizat două grafice, unul care prezintă atât valorile actuale, cât și cele previzionate ale variabilei dependente, iar celălalt distribuția față de 0 a diferenței dintre cele două valori. Valorile prezise sunt ilustrate cu culoarea albastră, iar cele actuale cu roșu.

Chart, histogram

Description automatically generated

*Figura 8*

Chart, scatter chart

Description automatically generated

*Figura 9*

Cu ajutorul graficelor din figurile anterioare (Figura 8 și Figura 9), putem observa o îmbunătățire a calității predicției față de modelul liniar, deoarece diferențele dintre valorile prezise și cele actuale sunt mai apropiate de 0.

În scopul optimizării modelului arborelui de decizie simplu, am construi modele alternative pentru a determina combinația ideală dintre *minsplit* și *maxdepth*. Combinația identificată a fost *minsplit* = 11, *maxdepth* = 20 și *cp* = 0,01. Deși am obținut o combinație a modelului optim, observăm faptul că valoarea RMSE-ului este aceeași cu cea rezultată folosind arborele implicit.

Text

Description automatically generated

*Figura 10*

Graphical user interface, text, application

Description automatically generated

*Figura 11*

**3. Arbore de decizie folosind bagging**

În încercarea de a îmbunătăți performanța modelului bazat pe arborii de decizie, am folosit metoda bagging. Prin urmare, am construit 3 modele diferite, folosind valoarea implicită pentru numărul de *bags,* o altă valoare descoperită ca fiind optimă (15, comparativ cu cea implicită de 25), și un model de bagging determinat prin *cross-validation*.

Putem compara performanțele modelelor create raportându-ne la valorile erorii de predicție (RMSE). Am ilustrat în imaginile următoare valorile erorii menționate, de unde deducem faptul că modelul cel mai performant este cel construit prin *cross-validation.*

A picture containing text

Description automatically generated

A picture containing graphical user interface

Description automatically generated

Graphical user interface, application

Description automatically generated with medium confidence

Chart

Description automatically generatedDe asemenea, modelul CVBagged determină importanța variabilelor, unde se poate observa că *overall, potential* și *wage\_eur* sunt considerate cele mai relevante. Putem remarca faptul că, spre deosebire de modelul liniar, variabila *age* nu este considerată la fel de importantă.

*Figura 12*

Pentru a evalua performanța modelului, am realizat două grafice, unul care prezintă atât valorile actuale, cât și cele previzionate ale variabilei dependente, iar celălalt distribuția față de 0 a diferenței dintre cele două valori. Valorile prezise sunt ilustrate cu culoarea albastră, iar cele actuale cu roșu.

Chart, histogram

Description automatically generated

*Figura 13*

Chart, scatter chart

Description automatically generated

*Figura 14*

Putem remarca faptul că acest model nu aduce o îmbunătățire vizibilă la nivelul graficului față de modelul creat cu arborele simplu.

**4. Arbori de decizie folosind random forests**

Am continuat cu construirea modelului de decizie bazat pe random forests. Inițial, s-a considerat modelul construit cu valorile implicite pentru numărul de arbori (500), mărimea *bag-*ului (0.632) și *mtry* (m/3 => 48/3 = 16). Performanța acestui model este reflectată prin valoarea erorii de predicție (RMSE), care este mai mică cu ~500.000 de euro comparativ cu modelul anterior (CVBagged).



Text

Description automatically generated Am căutat să optimizăm valorile pentru *mtry, sample\_size* și *node\_size.* Valorile rezultate ca fiind optime sunt: *mtry* = 48 (adică toate variabilele din setul de date), *sample\_size* = .8 (adică mărimea *bag*-ului fiind 80% din setul de antrenament), *node\_size* = 3 (adică numărul minim de observații dintr-un nod).

*Figura 15*

Text

Description automatically generatedLuând în considerare caracterul volatil al performanței modelului arborilor de decizie la mici schimbări ale setului de antrenament, am considerat necesară construirea modelului optim de 100 de ori.

*Figura 16*

Astfel, media erorii de predicție (RMSE) pe setul de antrenament a celor 100 de modele este următoarea:

Text

Description automatically generated with medium confidence

Performanța acestui model este reflectată prin valoarea erorii de predicție (RMSE) pe setul de testare, care este mai mică cu ~250.000 de euro comparativ cu modelul anterior (simpleRandomForest).



De asemenea, putem vizualiza pe următorul grafic importanța variabilelor conform modelului analizat. Observăm faptul că primele 2 variabile sunt aceleași precum în cadrul modelului CVBagged, și anume *overall* și *potential.* Pe de altă parte, avem pe locul 3 variabila *age*, regăsită în cadrul modelului CVBagged doar pe locul 6.

Graphical user interface, application

Description automatically generated

*Figura 17*

Pentru a evalua performanța modelului, am realizat două grafice, unul care prezintă atât valorile actuale, cât și cele previzionate ale variabilei dependente, iar celălalt distribuția față de 0 a diferenței dintre cele două valori. Valorile prezise sunt ilustrate cu culoarea albastră, iar cele actuale cu roșu.

Chart, histogram

Description automatically generated

*Figura 18*

Chart, histogram, scatter chart

Description automatically generated

*Figura 19*

**Concluzie**

În concluzie, am construit 4 modele de predicție a valorii jucătorilor, și anume modelul liniar, arbore de decizie simplu, arbore de decizie folosind bagging cu *cross-validation* și random forests optimizat pentru setul nostru de date. Ca rezultat, cel mai bun model este ultimul menționat anterior, deoarece are cea mai mică valoare a RMSE-ului. De asemenea, am obținut o scădere de la 2.383.658 la 760.422,2 euro, adică o reducere de 68% a erorii.

În imaginea de mai jos, se poate observa îmbunătățirea treptată a performanței modelului utilizând cele 4 tehnici menționate anterior prin reducerea erorii de predicție a valorii jucătorilor, deoarece diferențele dintre valorile previzionate și cele reale sunt mai apropiate de 0. Pe primul rând, putem citi de la stânga la dreapta modelul liniar, respectiv arborele de decizie simplu. Pe cel de-al doilea rând, tot de la stânga la dreapta se află arborele de decizie cu bagging cu *cross-validation*, respectiv arborele de decizie cu random forests optimizat.

Chart

Description automatically generated with medium confidence

*Figura 20*

**Bibliografie**

[1] https://www.statista.com/chart/16875/super-bowl-viewership-vs-world-cup-final, accesat la 28.05.2021

[2] https://www.statista.com/statistics/1118294/football-shirt-sales-by-club, accesat la 28.05.2021

[3] https://www.statista.com/statistics/271581/revenue-of-soccer-clubs-worldwide, accesat la 28.05.2021