Predicció de resultats dels gols i resultats de futbol amb Machine Learning

Barcelona, 23/05/2023

Borja Garcia-Mila
Data Science student of it Academy
Barcelona, Espanya
b.garmillo@gmail.com

Resum— Intentar predir el futbol o qualsevol esport a partir de dades estadístiques sempre ha estat un repte per una part de la societat i un misteri saber fins a quin punt pot ser previsible un resultat o no. Aquest treball és un intent més a partir d'unes variables subjectives per l'autor d'entrenar un model de Machine Learning. És un camp amb molt recorregut així que aquesta feina pot ser un punt de partida. Per fer-ho s'ha provat diferents models de regressió i classificació per veure quin aportava les millors mètriques i resultats.

Paraules claus— Python, Anàlisi, Machine Learning, Regressió, Classificació, PCA

1. INTRODUCCIÓ

Sempre s'ha defensat que en l'esport pot passar de tot ja que hem vist molts casos en que un equip de futbol o un esportista, contra pronòstic, ha canviat el guió de la història. Però fins a quin punt és imprevisible l'esport?

Més enllà de les cases d'apostes, en els darrers anys, diferents estudis han posat en dubte aquesta teoria. Gràcies a la investigació en la ciència de dades i intel·ligència artificial s'han començat a crear models en que poden predir molt més que els resultats finals d'un partit de futbol. De fet, la majoria dels equips de màxim nivell utilitzen molts d'aquests models pel seu propi benefici, com per exemple, veure l'estat físic dels seus jugadors.
L'objectiu d'aquest treball és fer un estudi de com certes variables subjectives, poden ajudar a predir el número de gols que hi haurà en un partit o si hi haurà guanyador i quin amb l'ajuda dels models supervisats de regressió i classificació de Machine Learning.

2. ESTAT D'ART[2, 3, 4]

Des de fa anys que s'intenta predir els resultats esportius, tant per les cases d'apostes com pels interessos dels propis clubs. Tot depèn de les variables que s'utilitzen. Algunes de les tècniques més habituals son: Anàlisis de dades històriques, models basats en classificació, models de classificació jeràrquica, us de dades en temps real i models d'aprenentatge per reforç.

Tot i que la intel·ligència artificial cada cop s'usa més en l'esport, intentar predir el resultat d'un partit de futbol, o qualsevol altre esport, no deixa de ser un repte ja que hi poden arribar a influir moltes variables, algunes d'elles molt poc controlables i per això, és pràcticament impossible, a dia d'avui, aconseguir una predicció perfecte.

De totes maneres, més enllà de les cases d'apostes i els propis equips, cada cop hi ha més gent que ho intenta fer, cosa que facilita a més gent interessada en aquesta cerca de l'impossible.

3. METODOLOGIA

A) DATASET[1]

El conjunt de dades amb el qual s'ha iniciat a fer l'anàlisi i entrenament dels diversos

models de Machine Learning, ha estat creat a partir de dos data set inicial extrets de la pàgina web BDFutbol.com.

Les dades inicials d'aquests data set eren, les informatives del dia del partit i resultat d'aquest (amb número de gols) i, per altra banda, la classificació de cada equip per temporada. A partir del primer data set s'han creat una sèrie de columnes segons els darrers últims partits abans de la disputa

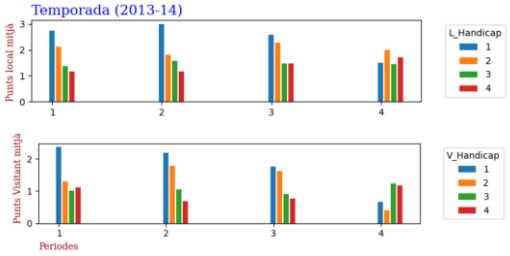

Fig. 1. Gràfic del dataset que mostra la mitja dels punts aconseguits segons el handicap de l'equip repartits ens els períodes de la temporada

d'aquest. A més a més, s'ha creat una altra variable, a partir del data set de la classificació en que s'ha fet una classificació de l'1 al 4 segons la classificació final dels equips en les darreres 10 temporades. En total hem començat amb un data set de 25 columnes i 26.525 files que quedaria reduït a 15.766 files ja que aquest estudi és a partir de la temporada 1980-81.

B) MODELS

Abans d'utilitzar els models he fet un estudi d'algunes de les variables, com es comportaven entre elles i he creat algunes visualitzacions d'interès de les columnes més interessants (període de temporada, handicap). També s'ha fet una neteja, ja que hi havia algun NaN i un estudi de la distribució de les variables per poder saber quin tipus de preprocessat aplicar a cada una. Com que teníem diversos targets (Gols locals, Gols visitants i resultat tipo quiniela) s'ha decidit fer dos tipus d'aprenentatge: supervisat de regressió per intentar predir els gols i de classificació per saber el resultat final del partit.

b.1) ML de regressió

En Machine Learning una de les llibreries més utilitzades és Scikit-learn que conté varis algorismes. Els models utilitzats han estat:

- LinearRegression
- KNeighborsRegression
- SVR

Per avaluar els resultats s'ha fet ús de les mètriques MAE i R2. En el cas del R2 s'ha de tenir en compte que els resultats de les prediccions es donen amb decimals per tant, hem ajustat les prediccions perquè ens donés un número enter. S

S'ha realitzat un RandomizedSearch per poder millorar els paràmetres prioritzant el temps. S'ha intentat millorar el model a través del PCA. Un cop trobat el millor model, s'ha procedit a aplicar-ho també en els gols visitants i finalment s'ha procedit a veure els gols per equip i si coincidia amb el que ha passat a la realitat.

b.2) ML de classificació.

Conscient que predir els gols és molt complicat s'ha creat un model de classificació per veure si guanyava l'equip local, visitant o cap. S'han creat 5 models comuns per veure quin és el que funcionava millor. Aquests han estat:

- KNeighborsClassifier
- DecisionTreeClassifier
- SVC
- RandomForestClassifier
- LogisticRegression

S'han buscat els millors paràmetres per cada un dels models. PCA no tenia sentit tornar-lo a fer ja que el procediment era el mateix. El resultat dels models amb híper-paràmetres ha estat similar utilitzant les mètriques per f1 micro (al no ser binaria la classificació). El millor model ha estat LogisticRegression amb el qual s'ha avaluat més a fons amb una

matriu de confusió i un report.

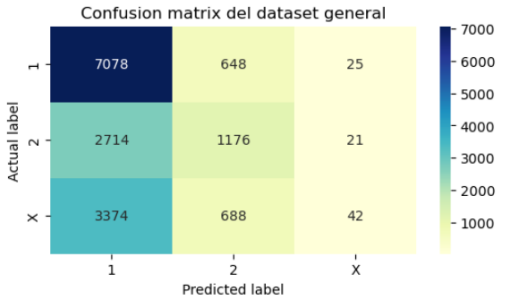

Fig. 2. Matriu de confusió del model Logistic Regression

Al tenir mostres molt descompensades en el target he utilitzat la tècnica de balanceig OverSampling per veure si millorava el resultat.

4. RESULTATS

Les dades aportades en la predicció ens ha donat com a millor valor de R^2 0.145 i MAE al voltant d'1. Els valors de R^2 van de 0 a 1 sent 1 que el model funciona perfectament. Per altra banda, MAE ens indica la distància mitja del valor absolut, és a dir de quan s'equivoca de mitja el model, per tant, pot anar de 0 a 10. Aplicar PCA o TruncatedSVD

```
MAE para LR_1: 0.967
R2 para LR_1: 0.124

MAE para KNR_1: 0.953
R2 para KNR_1: 0.145

MAE para SVM_1: 0.926
R2 para SVM_1: 0.107
```

Fig. 3. Resultats de les principals mètriques d'avaluació dels models de regressió en la predicció dels gols locals

(quan falta dada) no ha millorat els resultats. En l'aprenentatge de classificació els models obtenen una accuracy superior al 0,5 (És a dir millor que l'aleatorietat que seria un 33%). Tot i tenir les dades descompensades fer un balanceig no millora el resultat. He comprovat els resultats, ja que ja s'ha jugat la jornada: en regressió el model ha encertat 1 cop el nombre de gols per equip, 2 cops el nombre de gols de l'equip local i 4 cops l'equip visitant. Els resultats són millors que les mètriques i sorprèn el dels gols visitants ja que el model donava pitjor resultat que el del local. Pel que fa a la desviació del número de gols comprovem que en la majoria dels casos és similar al valor de MAE, aproximant-se a 1. Per altra banda, els resultats en classificació s'ajusten al model amb un 60% d'encerts.

Local	Visitant	Res	Pred d	Pred Up	Enc		Pred Down	Enc	
Betis	Rayo	3-1	1,6-1,1	2-1	N	S	1-1	N	S
Celta	Valencia	1-2	1,6-1,3	2-1	N	N	1-1	S	N
Villareal	Bilbao	5-1	1,8-0,9	2-1	N	S	1-0	N	N
Elche	Atlètico	1-0	0,8-1,8	0-2	N	N	0-1	N	N
Espanyol	Barça	2-4	1-1,9	1-2	N	N	1-1	N	N
R. Madrid	Getafe	1-0	2,4-0,9	2-1	N	N	2-0	N	S
Mallorca	Cadiz	2-1	1,7-0,7	2-1	S	S	1-0	N	N
Valladolid	Sevilla	0-3	1,2-0,8	1-1	N	N	1-0	N	N
Osasuna	Almeria	3-1	1,6-0,8	2-1	N	S	1-0	N	N
R. Societat	Girona	2-2	1,6-1,1	2-1	S	N	1-1	N	N
RESULTAT				1	2	4	0	1	2

Fig. 4. Comparativa dels gols predits amb el que ha passat realment amb el model de regressió

Local	Visitant	Res	Qui	Pred	Enc
Betis	Rayo	3-1	1	1	S
Celta	Valencia	1-2	2	1	N
Villareal	Bilbao	5-1	1	1	S
Elche	Atlètico	1-0	1	2	N
Espanyol	Barça	2-4	1	2	N
R. Madrid	Getafe	1-0	1	1	S
Mallorca	Cadiz	2-1	1	1	S
Valladolid	Sevilla	0-3	2	2	S
Osasuna	Almeria	3-1	1	1	S
R. Societat	Girona	2-2	X	1	N
RESULTAT					6

Fig. 5. Comparativa dels resultats del partit amb el que ha predit el model de classificació

5. CONCLUSIONS

Tot i que els valors reflectits en les mètriques de Machine Learning son baixos si que veiem certa aproximació, i per exemple, en aquesta jornada, ha donat millor resultat la realitat que les mètriques.

També veiem que les variables més interessants que he creat si tenen certa rellevància en els resultats (Períodes i Handicap) i que sovint son poc estudiades. En comparació amb altres estudis realitzats moltes variables son diferents i normalment s'utilitzen més que les utilitzades en aquest cas.

Algunes formes de millorar el projecte podria ser estudiar nous models i millor els paràmetres que podem aportar a veure si milloren les mètriques. Una altra cosa que em sembla interessant d'estudiar és veure com es comporten els models modificant la base de dades a només l'històric del mateix partit en altres temporades i que passa si treus els dos equips més estables amb

diferència. També es pot modificar les dades a veure que passa si ho canvies més o menys enllà dels últims 10 partits.

La conclusió és que el projecte em sembla interessant com a punt de partida per posar una mica de llum a la foscor però, com amant de l'esport, també te la seva part romàntica que tingui certa imprevisibilitat. A més a més hi ha certes variables molt difícils de preveure (errors arbitrals, lesions...).

6. Referències

1. BDFutbol : https://www.bdfutbol.com
2. ChatGPT
3. Wikiwand: https://www.wikiwand.com/es/Apuesta_deportiva
4. Big Data Sports: https://bigdatasports.media/2021/02/28/se-puede-predecir-el-futbol/