PREDICTIBILITY IN FOOTBALL

OUTLOOK

As occasional user of sports betting I have always wondered to what degree a football results can be predicted and if I can use these predictions to optimize both risk and gains for this kind of betting. The main idea is answering the following questions:

- How much **accurate** football results can be (1 X 2)?
- Can we identify **under or overpriced odds**?
- Can we **identify some niches** (by kind of match, team,etc) where the predictability/profitability are better?

STATE OF ART

There are several websites where you can check the % for every match of ending up in home win/draw/away win but they don´t neither say how these percentages are calculated nor offer different predicting models depending on the specific match or other variables and above all **they don´t show the most interesting matches in terms of potential earns**. By this analysis I will try to cover all these aspects

ENVIROMENT

It will be required to run every notebook or the one which contains the whole code from the next **repository from Github**.

It will be required the installation of software to run notebooks **(i.e. JupyterLab)**. One of the notebooks will run the installation of two needed packages: datapackage and tensorflow) **[NOTEBOOK 2]**

It will be required access to **Tableau** (to update the user report).

Access to the API **betsapi2** (user and password are included in the notebooks).

THE SOURCES

On one side I need an historical dataset with both results and odds for the last 10 seasons. This will be base to create the models to predict the results of the next coming round.

Furthermore, I need to identify the next matches' round and theirs associated odds. For these matches will predict their results.

For that purpose the next sources will be needed:

- Seasons 09-10 17-18
- Seasons 19 20 y 20 21
- Next round and associated odds.

DATA

The **creation of an historical database** will be needed and will be fed by the sources mentioned in the previous point. I will only keep the following variables:

- **Div:** League Division
- **Date**: Match Date (dd/mm/yy)
- **Time**: Match Time
- **HomeTeam**: Home Team
- **AwayTeam**: Away Team
- **FTHG**: Full Time Home Team Goals
- **FTAG**: Full Time Away Team Goals
- **FTR**: Full Time Result (H Home Win, D Draw, A Away Win)
- **HS**: Home Team Shots
- **AS**: Away Team Shots
- **HST:** Home Team Shots on Target
- **AST**: Away Team Shots on Target
- **B365H**: Bet365 home win odds
- **B365D**: Bet365 draw odds
- **B365A**: Bet365 away win odds

For the next matches' round I will keep these:

- **Div:** League Division
- **Date**: Match Date (dd/mm/yy)
- **Time**: Match Time
- **HomeTeam**: Home Team
- **AwayTeam**: Away Team
- **B365H**: Bet365 home win odds
- **B365D**: Bet365 draw odds
- **B365A**: Bet365 away win odds

METHODOLOGY

A) IDENTIFY THE NEXT ROUND : GAMES & ODDS [NOTEBOOK 3]

On one side is required identifying the games of the next round by the next API ,on the other hand we need their associated odds which will be retrieved by the following API. Both datasets have an unique id for each match which will be used to create a single merged table:

id	home_name	away_name
102540251	Levante	Cadiz
102540243	Celta Vigo	Real Betis
102540245	Eibar	Barcelona
102540249	Huesca	Valencia

id	1	X	2
102540251	1.727	4.100	4.200
102540243	3.200	3.800	2.050
102540245	6.000	5.000	1.444
102540249	1.615	4.200	5.000

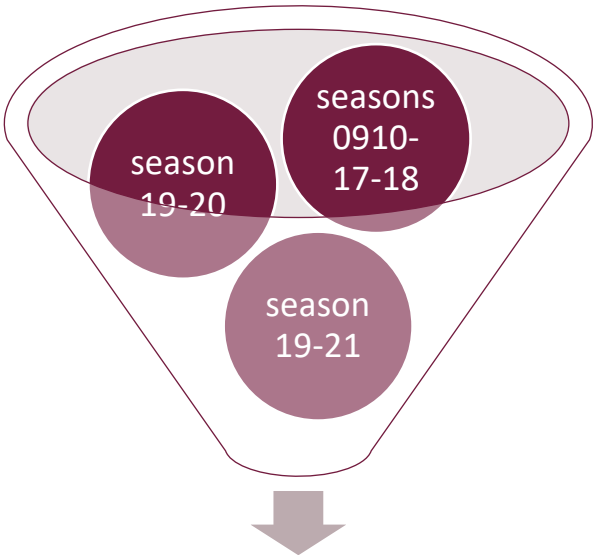
id	home_name	away_name	1	X	2
102540251	Levante	Cadiz	1.727	4.100	4.200
102540243	Celta Vigo	Real Betis	3.200	3.800	2.050
102540245	Eibar	Barcelona	6.000	5.000	1.444
102540249	Huesca	Valencia	1.615	4.200	5.000

IMPORTANT: due to the season is already over we will use the last round of this season as it was the next one.

B) CREATION OF AN HISTORICAL DATABASE [NOTEBOOK 4]

B.1) Data imports

The data is located in different blocks of season so it will be necessary to gather them to have an unique data source:



	Div	Date	HomeTeam	AwayTeam	FTHG	FTAG	FTR	HS	AS	HST	AST	B365H	B365D	B365A	season
0	SP1	2018-08-17	Betis	Levante	0	3	A	22	6	8	4	1.66	4	5	1819
1	SP1	2018-08-17	Girona	Valladolid	0	0	D	13	2	1	1	1.75	3.6	5	1819
2	SP1	2018-08-18	Barcelona	Alaves	3	0	H	25	3	9	0	1.11	10	21	1819
3	SP1	2018-08-18	Celta	Espanol	1	1	D	12	14	2	5	1.85	3.5	4.5	1819
4	SP1	2018-08-18	Villarreal	Sociedad	1	2	A	16	8	7	4	2.04	3.4	3.8	1819

The final dataset will be made of ~ 4500 games and during this step a 'season' variable will be added to identify the season of each game.

B.2) Dataset by team/season creation.

The **available variables** can be seen as **very limited** to expect to have reliable predictions.

Therefore, will be needed **having the number of the round and points of every team just before of starting the next round**.

That´s why we need to calculate these variables for each team in each round by identifying they were playing as local or visitor and depending on the result assign them the points..

season	Team	Date	Home/Away	match	Div	FTHG	FTAG	FTR	HS	AS	HST	AST	B365H	B365D	B365A	points
0910	Almeria	2009-08-30	H	1	SP1	0	0	D	20	7	5	1	2.1	3.3	3.5	1
0910	Almeria	2009-09-13	A	2	SP1	1	0	H	16	7	4	0	2.38	3.25	3	0
0910	Almeria	2009-09-20	H	3	SP1	1	0	H	11	23	3	11	2.5	3.25	2.8	3
0910	Almeria	2009-09-23	A	4	SP1	2	2	D	24	12	9	7	1.44	4.33	7	1
0910	Almeria	2009-09-27	H	5	SP1	2	2	D	13	13	4	6	2.25	3.25	3.2	1
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2021	Villarreal	2021-04-25	H	33	SP1	1	2	A	10	15	4	5	4.5	4.2	1.66	0
2021	Villarreal	2021-05-02	H	34	SP1	1	0	H	4	15	3	3	2.1	2.9	4.1	3
2021	Villarreal	2021-05-09	H	35	SP1	2	4	A	16	11	7	6	1.95	3.6	3.8	0
2021	Villarreal	2021-05-13	A	36	SP1	0	2	A	11	9	2	3	3.8	3.5	1.95	3
2021	Villarreal	2021-05-16	H	37	SP1	4	0	H	12	17	5	5	2.37	3.4	2.87	3

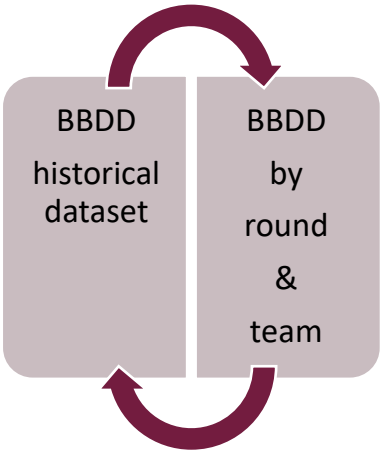
Subsequently, some calculations have to be made: accumulated shots for/against, acc goals for/against, acc points for/against till the previous round to the one that will be estimated for every team in every round.

goals_for_acc	goals_against_acc	shots_for_acc	shots_against_acc	shots_target_for_acc	shots_target_against_acc	points_acc
0	0	0	0	0	0	0
0	0	20	7	5	1	1
0	1	27	23	5	5	1
1	1	38	46	8	16	4
3	3	50	70	15	25	5
...	...	...	...	...	...	...
49	36	349	327	139	103	49
50	38	359	342	143	108	49
51	38	363	357	146	111	52
53	42	379	368	153	117	52
55	42	388	379	156	119	55

Once the accumulated points by team&round have been calculated the calculation of the classification is needed. It will calculated conversely (20th is the leader and 1 is the bottom team) in favor of interpretability to keep a positive relation with the other variables

season	Date	match	Team	Div	points_acc	ranking
0910	2009-08-30	1	Almeria	SP1	0	1
0910	2009-09-13	2	Almeria	SP1	1	9
0910	2009-09-20	3	Almeria	SP1	1	3
0910	2009-09-23	4	Almeria	SP1	4	11

B.3) Merge data from the calculated data and from the original database.



The BBDDs of the football matches and the one created by team and day will be crossed. In this way I will have accumulated goals(for/against), points, shots for each team (for/against) and classification.

To enable the machine learning models to interpret data will be necessary to calculate the difference of every variable between the two teams of each match.

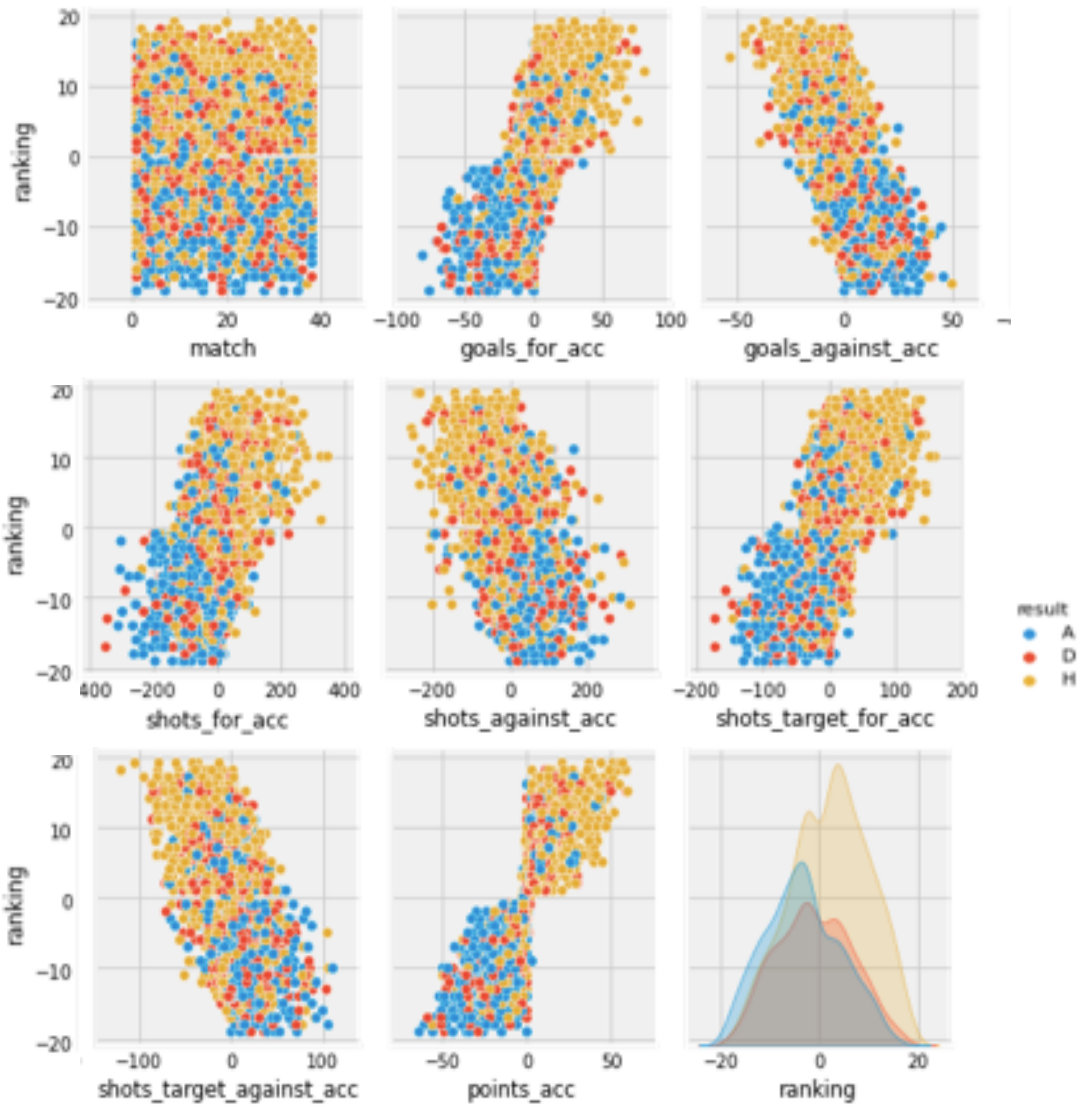
.

season	match	HomeTeam	AwayTeam	goals_for_acc	goals_against_acc	shots_for_acc	shots_against_acc
2021	37	Cadiz	Elche	4	-1	52	-47
2021	37	Getafe	Levante	-17	-11	-22	-140
2021	37	Sociedad	Valladolid	22	-14	66	-135
2021	37	Valencia	Eibar	18	5	-40	171
2021	37	Villarreal	Sevilla	3	13	-45	52

season	match	HomeTeam	AwayTeam	shots_target_for_acc	shots_target_against_acc	points_acc	ranking	result
2021	37	Cadiz	Elche	15	-15	13	7	A
2021	37	Getafe	Levante	-24	-52	-6	-3	H
2021	37	Sociedad	Valladolid	35	-49	25	13	H
2021	37	Valencia	Eibar	7	31	9	6	H
2021	37	Villarreal	Sevilla	19	2	-19	-2	H

Example: Getafe got to the day 37 with 17 goals less than Levante,
3 positions below since had 6 points less than Levante.

C) CREATED DATASET EXPLORATION [NOTEBOOK 5]



It is not surprising that when the ranking is positive (meaning that the local team classifies better than the visitor) and variables such as accumulated points, goals or shots are positive too, the local wins become more frequent.

Draws don´t follow a clear pattern so it may be assumed that will be the result more difficult to predict.

The day of the game seems that has not effect to predict the result.

D) PREPARING THE TRAIN/TEST DATASETS [NOTEBOOK 6]

D.1) Preprocessing data

The main idea is applying ML models to both the regular dataset and its scaled version to establish subsequently which model/kind of dataset choice is the optimum.

For that purpose, firstly we will apply LabelEncoder to the names of eams (HomeTeam & AwayTeam) and results (H-D-A).

The type of some variables will be modified:

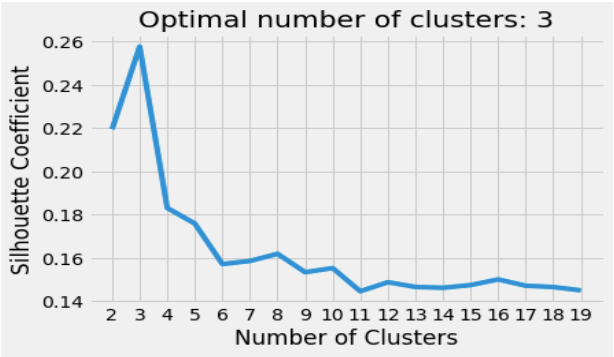
- categories = ['Div','season','HomeTeam','AwayTeam','result']
- floats = ['B365H','B365D','B365A']
- integers = ['season']

D.2) Splitting the dataset to train it

Before splitting the dataset all the variables will be defined as 'features' (Except the odds since are implicit predictions made by B365) and the result will be defined as 'target'.

Once the features and the target have been defined, the dataset will be split, setting 90%-10% train/test ratio. Then, a StandardScaler will be applied to the train dataset.

D.3) K-Means clustering

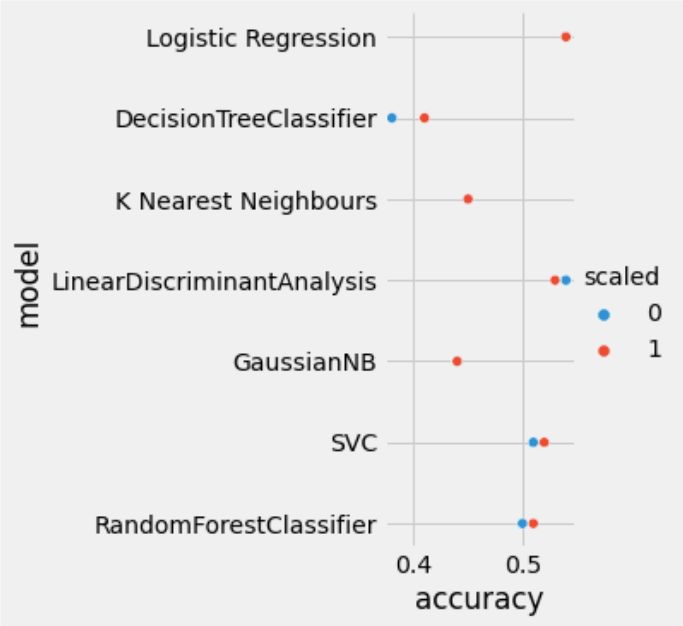


It should be expected that there are ´similar´ games which can have different levels of predictability. So, K-means will be applied to the train Scaled Dataset, to establish the optimum number of clusters.

Número de Clusters : 3

E) APPLYING DIFFERENT ML MODELS [NOTEBOOK 7]

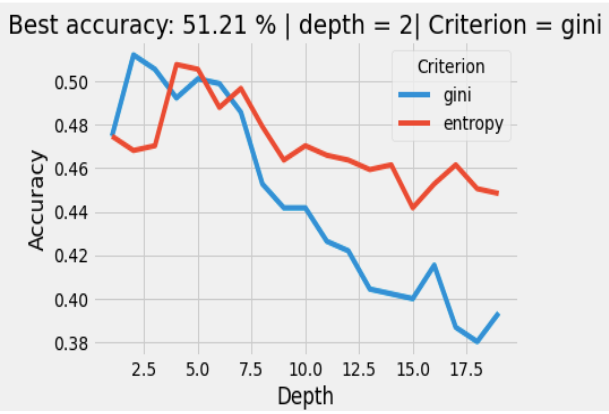
E.1) First approach with many models.

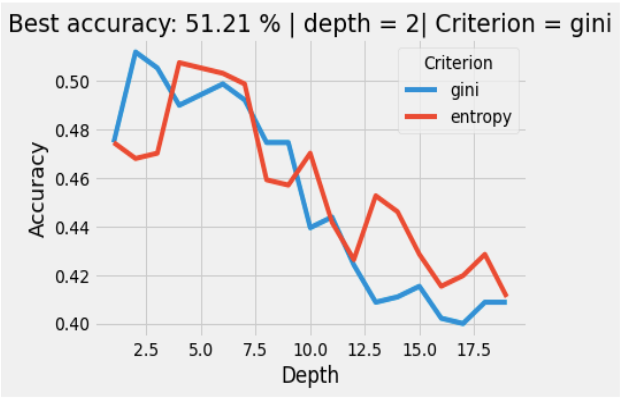


- 5 models were applied on the regular train dataset and on the scaled one and at best the accuracy is 50%.
- The differences between the regular and the scaled datasets were not significant.

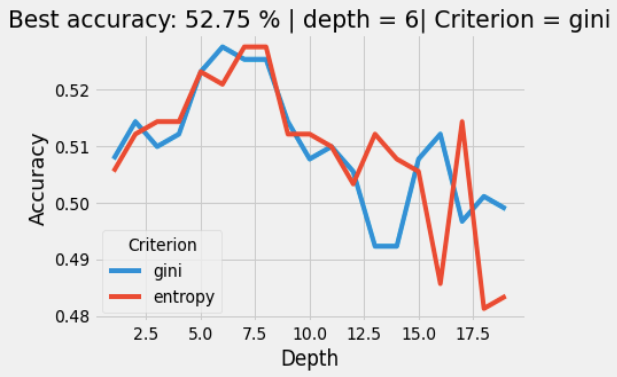
E.2) Then, the best parameters for each couple of datasets were looked for.

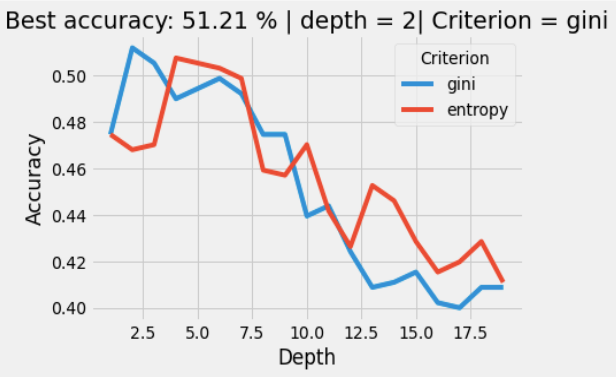
Decision Tree Classifier



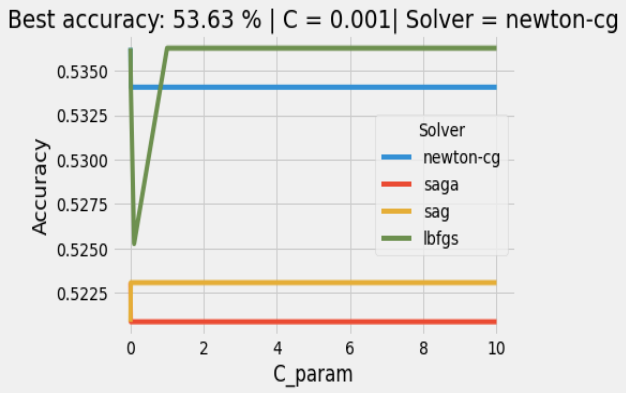


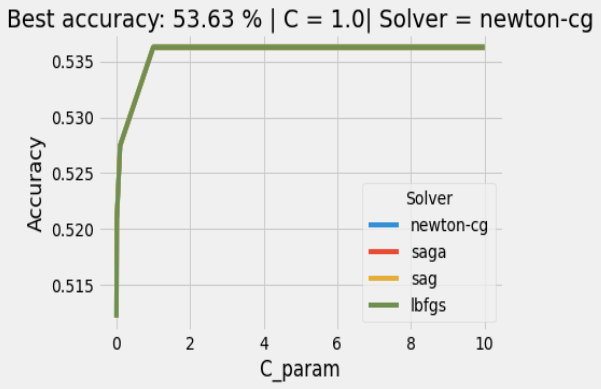
Random Forest Classifier



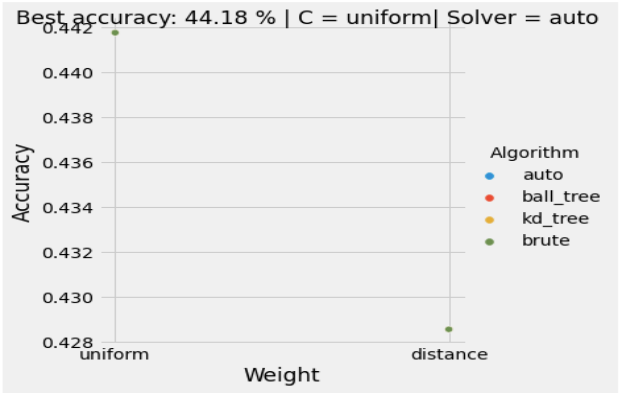


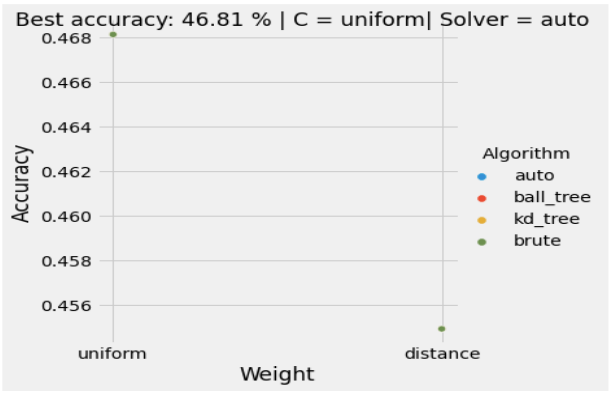
Logistic Regression



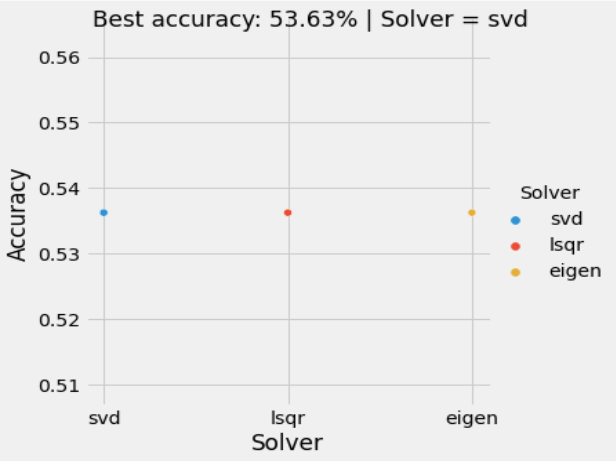


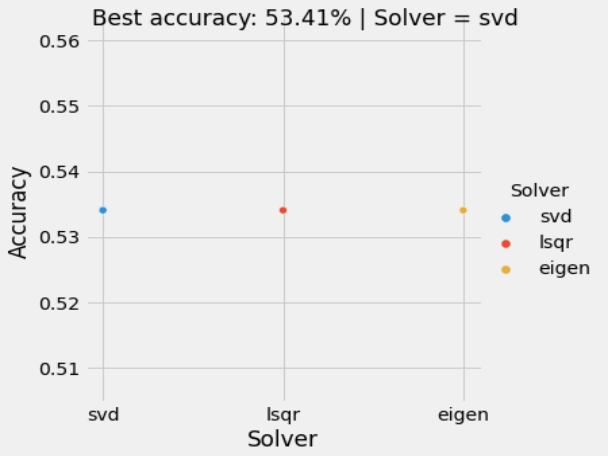
K-Neighbors Classifier



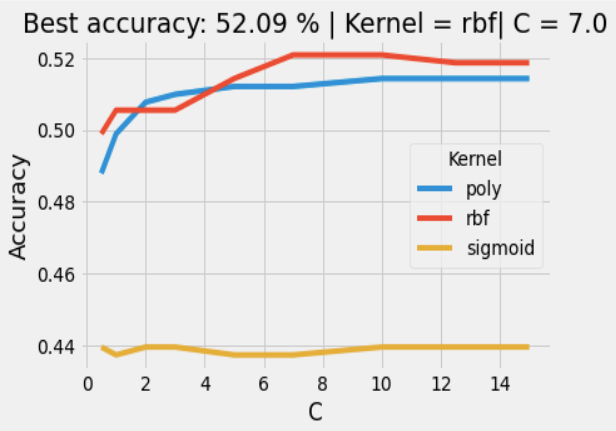


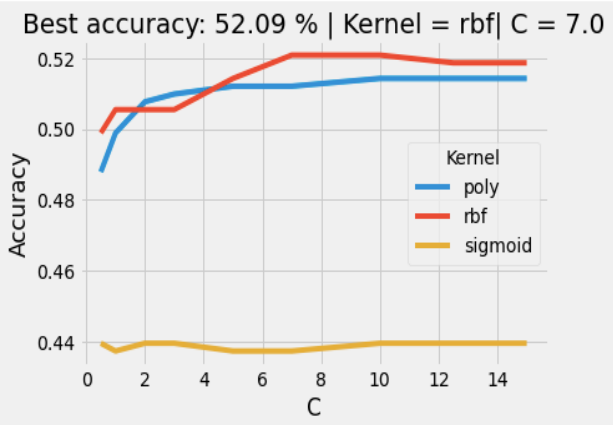
Linear Discriminant Analysis





SVC





E.3) Application of the models

Application of each model, with its best parameterization to optimize the accuracy, on both the 'X_test' and the whole dataset.

In this way there will be ,for each match, the prediction from every model and the percentages of the odds for local win, draw or visitor win (H,D,A).

RESULTS

F) FINAL DATASET GENERATION & CONCLUSIONS DRAWN [NOTEBOOK 8]

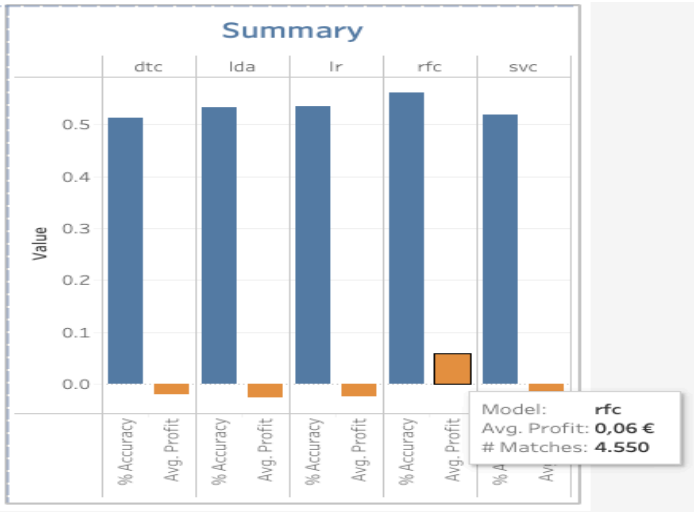
Whole database generation including every match, with their associated predictions for each potential result (H-D-A), to be used to draw the conclusiones by the Tableu report (link pendiente de linkar).

season	HomeTeam	AwayTeam	match_home	match_away	B365H	B365D	B365A	FTR	cluster
1819	Barcelona	Alaves	1	1	1.11	10	21	H	2
1516	Betis	Barcelona	36	36	21	9	1.13	A	0
2021	Eibar	Alaves	34	34	2,5	3,1	3	H	2

season	HomeTeam	AwayTeam	Prob_Home%	Prob_Draw%	Prob_Away%	Prediction	hit_result	paid_quota	Prize	Profit	Difficulty	Model
1819	Barcelona	Alaves	42,34	28,16	29,5	H	1	1.11	1.11	0.11	1	dtc
1516	Betis	Barcelona	10,81	21,34	67,85	A	1	1.13	1.13	0.13	1	rfc
2021	Eibar	Alaves	54,36	25,22	20,42	H	1	2,5	2,5	1,5	2	lr

It´s important to highlight that a good accuracy is the one that predicts the best. High accuracy with a negative/low payback might not be interesting while lower accuracy might be interesting if payback is positive.

Global Results



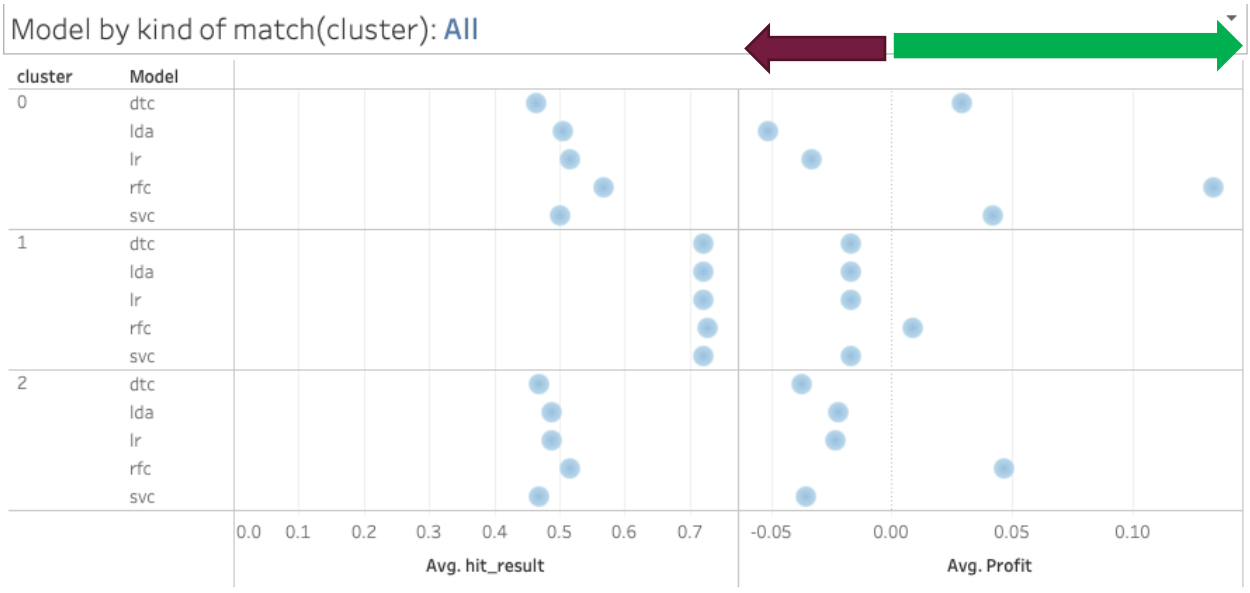
Every model is around 50% accuracy but only **Random Forest Classifier** produces 0.06€ benefit for each invested euro.

Result by type of game (Cluster)

When I started to create the dataset I decided to apply K-Means to know how many groups of games I could divide all the games in, to analyze later if any of these groups were more interesting in terms of accuracy/benefit.

The games from the cluster 1 are very interesting in terms of predictability since they have in average 70% accuracy, however the cluster 0 is more interesting in terms of profitability due to 3 out of 5 selected models give, in average, positive paybacks (dtc, rfc, svc).

The cluster 2 contains games as complicated to predict as the ones from the cluster 0 , however it gives, in average, lower paybacks so the ratio accuracy/benefit is less interesting.



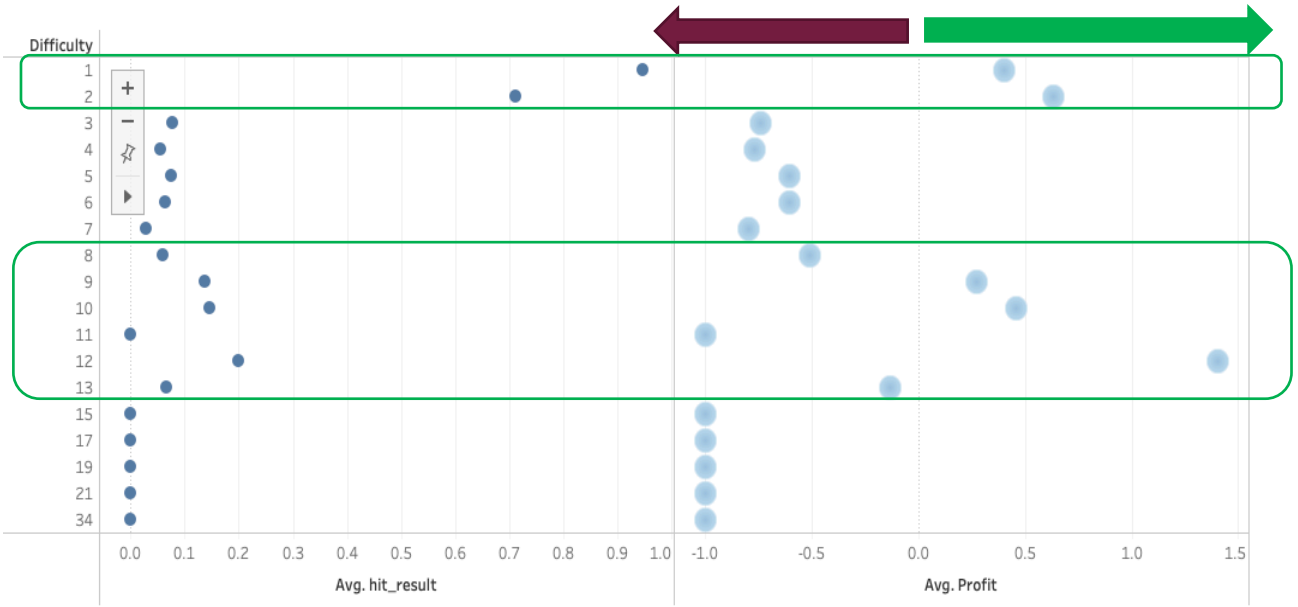
Result by odd amounts (Difficulty)

Other interesting approach would be analysing if it would be more interesting , long-term, to bet on low odds (more accurate) or high odds (less accurate). **Is better to bet , in general terms, with a risky or a conservative strategy?**

To answer this question, keeping the simplicity, the levels of difficulty have been defined as follows:

Odds that were the one related to the right result:

- From 1.01 to 1.99 → 1
- From 2.00 to 2.99 → 2
- ….
- From 34.00 to 34.00 → 34



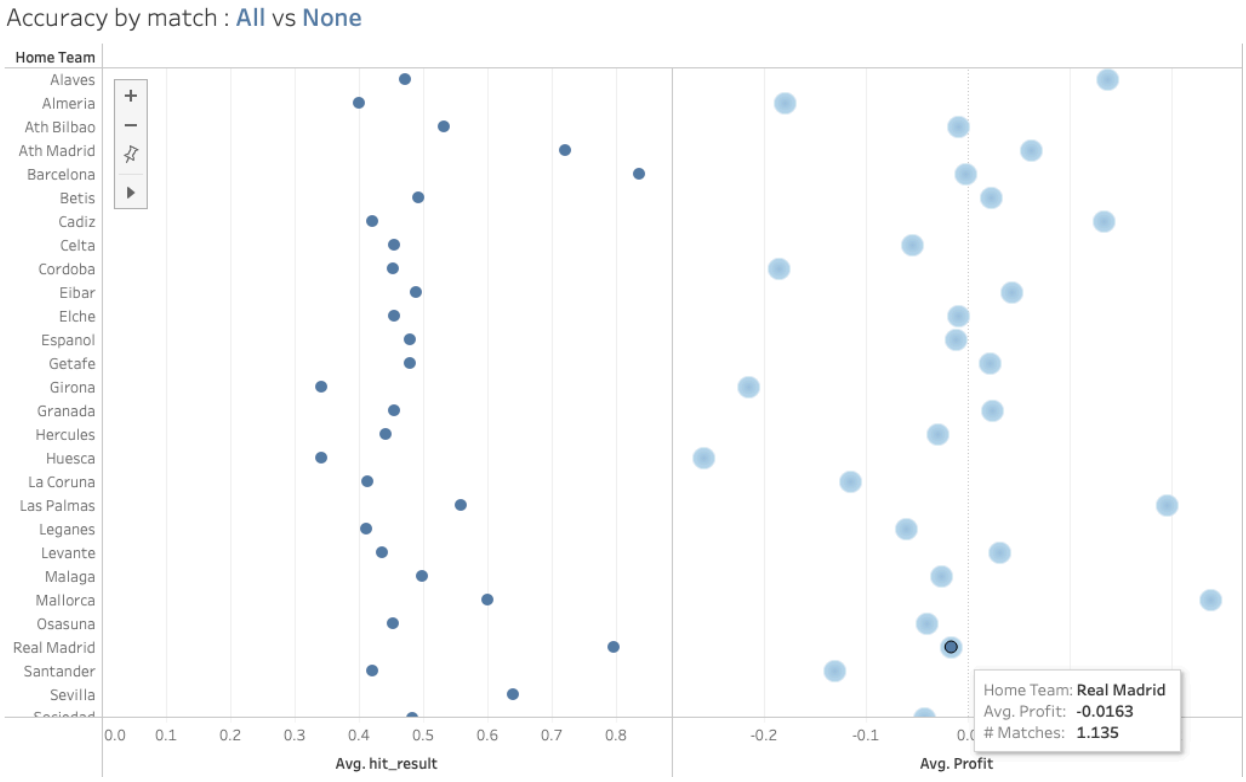
Checking the results it is more interesting to be ton odds lower than 2.99€ and in the case of betting riskly is more interesting to it in the interval

8-13€ since despite the accuracy is low the average profit is higher than the ones with odds between 3€ -7€ or 15€-34€.

Results by Home/Visitor teams

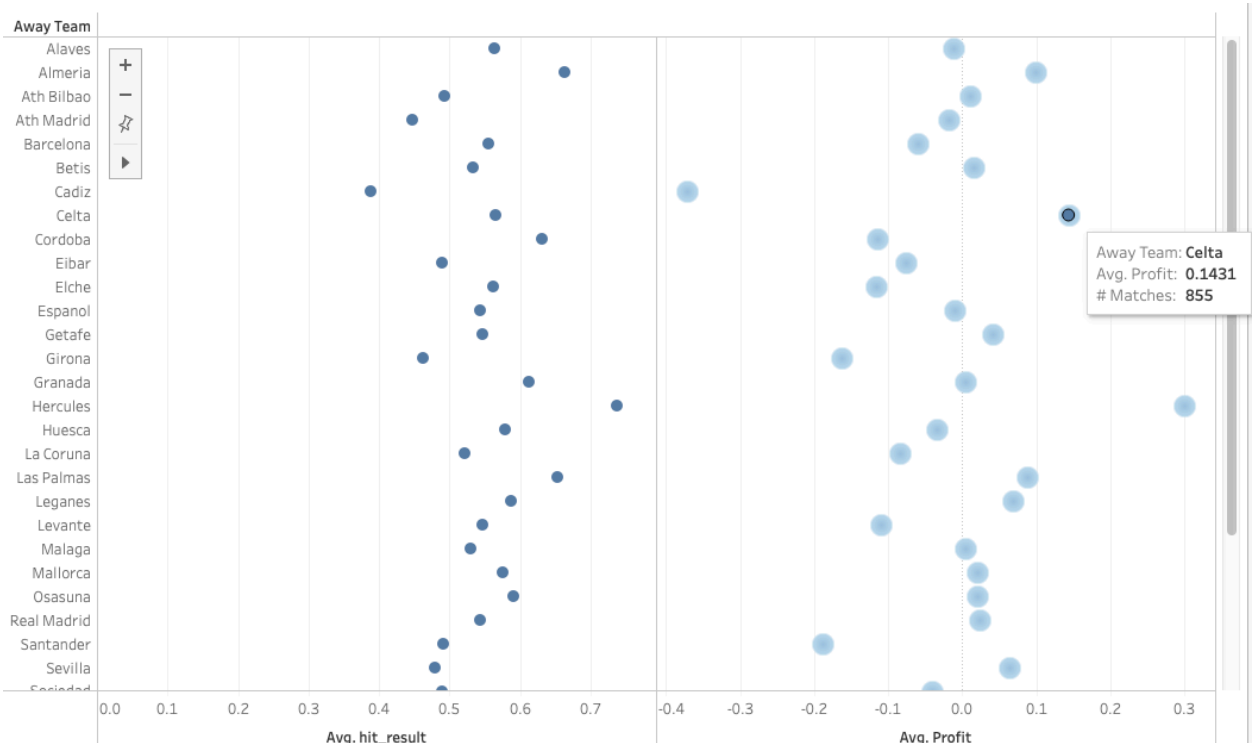
Even with no data to affirm it categorically, it might be intuitive to state that the most of Spanish football bets involve teams such as Madrid, Barca or Atlético. But are they the most interesting options?

As we can see with the example of Real Madrid, despite having an accuracy around 80% when playing at home, the average profit is negative, so other teams such as Alavés or Cádiz are more interesting bets when they play at home.



What about visitor teams?

Same pattern. Betting on Celta as a visitor is much more interesting than to doing it for the best teams since their results can be predicted correctly 56% of the time, leading us to a 15 cents average profit.



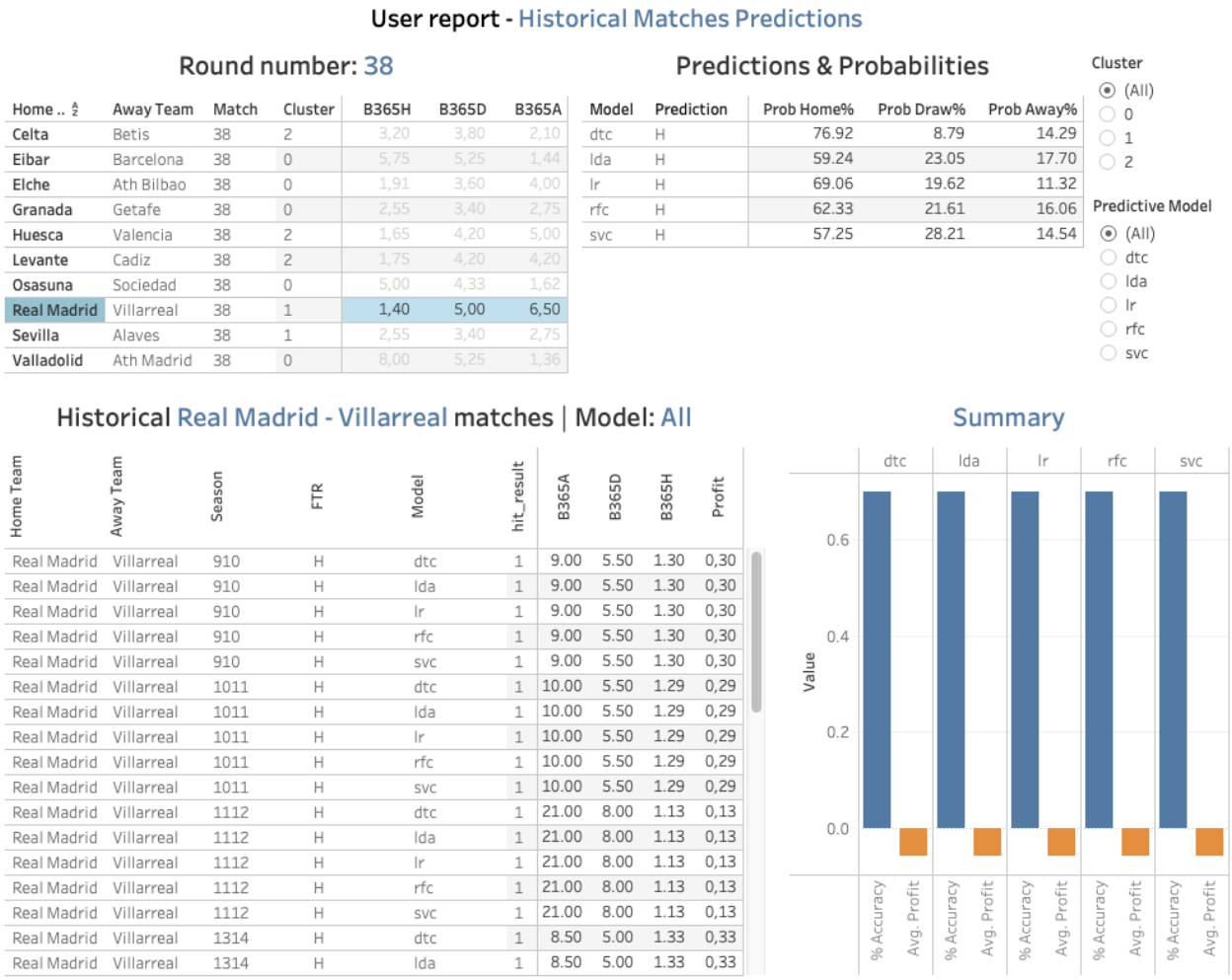
All these conclusions might be seen as a good reference overall, but next games come and we want to know which game to bet on. How do we make all this information available to any user? ..

G) USER REPORT GENERATION [NOTEBOOK 9]

The idea of the report is to offer all the information we have been so far But applied to the 10 gamess of the following league day [due to the Spanish league is over we will pretend than the last day is the coming day].

For this we will cross the matches, together with their quotas, obtained in The notebook 2 and we will apply all the processes (calculation of variables, K-Means clustering and ML models).

In this way we can see game by game which could be more Interesting. Shall we bet on Real Madrid?



User report - Historical Matches Predictions

Round number: 38

Home ..	Away Team	Match	Cluster	B365H	B365D	B365A
Celta	Betis	38	2	3,20	3,80	2,10
Eibar	Barcelona	38	0	5,75	5,25	1,44
Elche	Ath Bilbao	38	0	1,91	3,60	4,00
Granada	Getafe	38	0	2,55	3,40	2,75
Huesca	Valencia	38	2	1,65	4,20	5,00
Levante	Cadiz	38	2	1,75	4,20	4,20
Osasuna	Sociedad	38	0	5,00	4,33	1,62
Real Madrid	Villarreal	38	1	1,40	5,00	6,50
Sevilla	Alaves	38	1	2,55	3,40	2,75
Valladolid	Ath Madrid	38	0	8,00	5,25	1,36

Predictions & Probabilities

Model	Prediction	Prob Home%	Prob Draw%	Prob Away%
dtc	H	76.92	8.79	14.29
lda	H	59.24	23.05	17.70
lr	H	69.06	19.62	11.32
rfc	H	62.33	21.61	16.06
svc	H	57.25	28.21	14.54

Cluster
(All)
0
1
2

Predictive Model
(All)
dtc
lda
lr
rfc
svc

Historical Real Madrid - Villarreal matches | Model: All

Home Team	Away Team	Season	FTR	Model	hit_result	B365A	B365D	B365H	Profit
Real Madrid	Villarreal	910	H	dtc	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	910	H	lda	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	910	H	lr	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	910	H	rfc	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	910	H	svc	1	9.00	5.50	1.30	0,30
Real Madrid	Villarreal	1011	H	dtc	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1011	H	lda	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1011	H	lr	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1011	H	rfc	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1011	H	svc	1	10.00	5.50	1.29	0,29
Real Madrid	Villarreal	1112	H	dtc	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1112	H	lda	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1112	H	lr	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1112	H	rfc	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1112	H	svc	1	21.00	8.00	1.13	0,13
Real Madrid	Villarreal	1314	H	dtc	1	8.50	5.00	1.33	0,33
Real Madrid	Villarreal	1314	H	lda	1	8.50	5.00	1.33	0,33

Summary

What about Celta – Betis?

It would seem a game to avoid since accuracy of this games have been very low along their history. No model reaches 30% success rate and on average above 40 cents are lost per invested euro. To sum up these games is too complicated to guess in proportion to its odds.

User report - Historical Matches Predictions

Round number: 38

Home ..	Away Team	Match	Cluster	B365H	B365D	B365A
Celta	Betis	38	2	3,20	3,80	2,10
Eibar	Barcelona	38	0	5,75	5,25	1,44
Elche	Ath Bilbao	38	0	1,91	3,60	4,00
Granada	Getafe	38	0	2,55	3,40	2,75
Huesca	Valencia	38	2	1,65	4,20	5,00
Levante	Cadiz	38	2	1,75	4,20	4,20
Osasuna	Sociedad	38	0	5,00	4,33	1,62
Real Madrid	Villarreal	38	1	1,40	5,00	6,50
Sevilla	Alaves	38	1	2,55	3,40	2,75
Valladolid	Ath Madrid	38	0	8,00	5,25	1,36

Predictions & Probabilities

Model	Prediction	Prob Home%	Prob Draw%	Prob Away%
dtc	H	56.15	24.36	19.49
lda	H	42.98	28.57	28.45
lr	H	44.38	29.42	26.19
rfc	H	38.54	27.98	33.48
svc	H	35.90	27.06	37.04

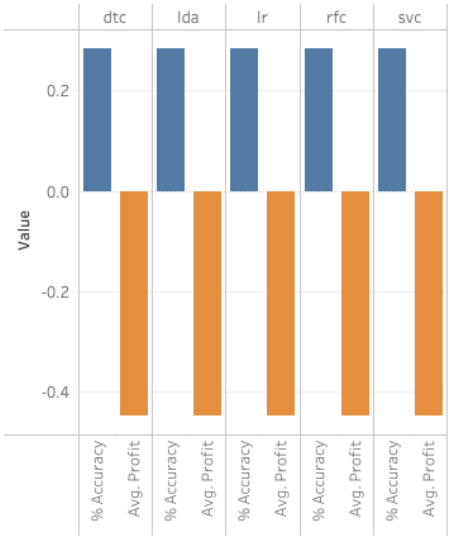
Cluster
(All)
0
1
2

Predictive Model
(All)
dtc
lda
lr
rfc
svc

Historical Celta - Betis matches | Model: All

Home Team	Away Team	Season	FTR	Model	hit_result	B365A	B365D	B365H	Profit
Celta	Betis	1213	A	dtc	0	3.500	3.400	2.050	-1,00
Celta	Betis	1213	A	lda	0	3.500	3.400	2.050	-1,00
Celta	Betis	1213	A	lr	0	3.500	3.400	2.050	-1,00
Celta	Betis	1213	A	rfc	0	3.500	3.400	2.050	-1,00
Celta	Betis	1213	A	svc	0	3.500	3.400	2.050	-1,00
Celta	Betis	1314	H	dtc	1	3.400	3.400	2.100	1,10
Celta	Betis	1314	H	lda	1	3.400	3.400	2.100	1,10
Celta	Betis	1314	H	lr	1	3.400	3.400	2.100	1,10
Celta	Betis	1314	H	rfc	1	3.400	3.400	2.100	1,10
Celta	Betis	1314	H	svc	1	3.400	3.400	2.100	1,10
Celta	Betis	1516	D	dtc	0	5.500	4.200	1.600	-1,00
Celta	Betis	1516	D	lda	0	5.500	4.200	1.600	-1,00
Celta	Betis	1516	D	lr	0	5.500	4.200	1.600	-1,00
Celta	Betis	1516	D	rfc	0	5.500	4.200	1.600	-1,00
Celta	Betis	1516	D	svc	0	5.500	4.200	1.600	-1,00
Celta	Betis	1617	A	dtc	0	3.750	3.400	2.050	-1,00
Celta	Betis	1617	A	lda	0	3.750	3.400	2.050	-1,00

Summary



FINAL COMMENTS

Generally speaking sports betting ,in the context of the Spanish soccer league, and only for 1-X-2 bets, and considering only the BET365 odds, l**eads to losses.**

However, and by analyzing the 10-year history, **we can identify kind of games that with a better ratio 'predictability / profitability'.**

The use of several models helps to optimize the decision since different models present different results for the same kind of games/bets.