



iscte

INSTITUTO
UNIVERSITÁRIO
DE LISBOA

PROJETO APLICADO EM CIÊNCIA DE DADOS I MODELING & EVALUATION

LICENCIATURA EM CIÊNCIA DE DADOS

Base de Dados ATP – Brasil

Grupo 5: nº 103303, nº 110451, nº 104716, nº 99239

Docentes: Diana Aldea Mendes e Sérgio Moro

24 de maio de 2023

Alterações ao trabalho



Adição de “feature importance” para cada modelo;



Adição de cross - validation com 10 folds para cada modelo;



Criação da variável “DifHands”



Novos modelos com variáveis diferentes.



Utilização de 10 folds para todos os modelos

Performance dos modelos com o conjunto de variáveis nº 1

Variáveis usadas	Modelos	Accuracy	Precision	Recall	F1-score	Support	Mean ROC
DifNumberWins	Modelo 1- XGBoost	0.61	0.66	0.84	0.74	1662	0.53
DifRank			0.32	0.15	0.20	840	
DifAge	Modelo 2- DecisionTree	0.54	0.66	0.64	0.65	1657	-
DifHeight			0.33	0.34	0.33	845	
Ground-clay	Modelo 3 - GradientBoosting	0.66	0.66	0.99	0.80	1657	-
Ground-hard			0.43	0.01	0.02	845	
Prize	Modelo 4- AdaBoost	0.66	0.66	0.98	0.79	1657	-
			0.34	0.02	0.04	845	

Cross Validation

com Random forest:

- 5 fold: 0.624
- 10 fold: 0.622

Observações:

- Entre os 4 modelos, os que têm uma melhor performance são os dois últimos;

Performance dos modelos com o conjunto de variáveis nº 2

Variáveis usadas
DifNumberWins
DifRank
DifAge
DifHeight
L_OR_R_Left-Handed
L_OR_R_Right-Handed
L_OR_R_Opponent_Left-Handed
L_OR_R_Opponent_Right-Handed
Ground_Clay
Ground_Hard

Modelos	Accuracy	Precision	Recall	F1-score	Support	Mean ROC
Modelo 1 - XGBoost	0.60	0.66 0.32	0.83 0.16	0.73 0.22	1657 845	0.53
Modelo 2 - DecisionTree	0.54	0.65 0.35	0.64 0.36	0.65 0.35	1635 867	-
Modelo 3 - GradientBoosting	0.65	0.65 0.39	0.99 0.01	0.79 0.02	1635 867	-
Modelo 4 - AdaBoost	0.65	0.65 0.31	0.99 0.00	0.79 0.01	1635 867	-

Cross Validation com Random forest:

- 5 fold: 0.606
- 10 fold: 0.609

Observações:

- O melhor modelo é o 3, apesar de ter uma performance mais baixa do que o modelo anterior.

Performance dos modelos com o conjunto de variáveis nº 3

Variáveis usadas
DifNumberWins
DifRank
DifAge
DifHeight
Prize
L_OR_R_Left-Handed
L_OR_R_Right-Handed
L_OR_R_Opponent_Left-Handed
L_OR_R_Opponent_Right-Handed
Ground_Clay
Ground_Hard

Modelos	Accuracy	Precision	Recall	F1-score	Support	Mean ROC
Modelo 1- XGBoost	0.61	0.66 0.36	0.84 0.18	0.74 0.24	1635 867	0.52
Modelo 2 - DecisionTree	0.54	0.66 0.32	0.64 0.35	0.65 0.34	1665 837	-
Modelo 3 - GradientBoosting	0.67	0.67 0.00	1.00 0.00	0.80 0.00	1635 867	-
Modelo 4- AdaBoost	0.66	0.67 0.43	0.99 0.01	0.80 0.03	1665 837	-

Cross Validation com Random forest:

- 5 fold: 0.618
- 10 fold: 0.628

Observações:

- O melhor modelo é o 4;
- O modelo 3 não prevê sets de 3.

Performance dos modelos com o conjunto de variáveis nº 4

Variáveis usadas	Modelos	Accuracy	Precision	Recall	F1-score	Support	Mean ROC
DifNumberWins	Modelo 1- DecisionTree	0.57	0.68 0.37	0.66 0.39	0.67 0.38	1667 835	-
DifRank							
DifAge	Modelo 2 - GradientBoosting	0.66	0.67 0.37	0.99 0.01	0.80 0.02	1667 835	0.53
DifHeight							
Prize	Modelo 3- AdaBoost	0.66	0.67 0.37	0.98 0.02	0.79 0.04	1667 835	0.52
DifHands							
	Modelo 4 – XGBoost	0.62	0.67 0.35	0.84 0.17	0.75 0.23	1667 835	0.52

Cross Validation com Random forest:

- 5 fold: 0.620
- 10 fold: 0.618

Observações:

- Os valores para as métricas dos 4 modelos são muito parecidos;
- O que se destaca (por pouco) é o modelo 2.

Performance dos modelos com o conjunto de variáveis nº 5

Variáveis usadas	Modelos	Accuracy	Precision	Recall	F1-score	Support	Mean ROC
DifNumberWins	Modelo 1- DecisionTree	0.55	0.67	0.65	0.66	1679	-
DifRank			0.34	0.36	0.35	823	
DifAge	Modelo 2 – GradientBoosting	0.66	0.67	0.99	0.80	1679	0.54
DifHeight			0.26	0.01	0.02	823	
Prize	Modelo 3- AdaBoost	0.66	0.67	0.98	0.80	1679	0.55
DifHands			0.33	0.02	0.04	823	
Ground_Clay	Modelo 4 – XGBoost	0.61	0.67	0.83	0.74	1679	0.51
Ground_Hard			0.31	0.16	0.21	823	

Cross Validation

com Random forest:

- 5 fold: 0.621
- 10 fold: 0.616

Observações:

- O modelo 3 destaca-se, não só pela accuracy mas também pela precision na previsão de sets de 2 e de 3.

Performance dos modelos com o conjunto de variáveis nº 6

Variáveis usadas
DifNumberWins
DifRank
DifAge
Prize

Modelos	Accuracy	Precision	Recall	F1-score	Support	Mean ROC
Modelo 1- DecisionTree	0.54	0.66 0.34	0.64 0.36	0.65 0.35	1652 850	-
Modelo 2 – GradientBoosting	0.65	0.66 0.29	0.99 0.01	0.79 0.01	1652 850	0.52
Modelo 3- AdaBoost	0.66	0.66 0.62	0.99 0.02	0.80 0.03	1652 850	0.53
Modelo 4 – XGBoost	0.61	0.66 0.33	0.84 0.16	0.74 0.21	1652 850	0.50

Cross Validation com Random forest:

- 5 fold: 0.604
- 10 fold: 0.608

Observações:

- O modelo 3 é o melhor modelo até agora, sendo que tem a melhor accuracy e previsão de sets de 2 e de 3;

Considerações finais

- Os melhores modelos utilizam o algoritmo AdaBoost ou GradientBoost (apesar de haver exceções);
- Apesar de maioria dos modelos ter uma accuracy que vai de 50% a 70%, a melhor accuracy conseguida: 67%;
- O melhor modelo em termos de previsão de sets de 2 e de 3 foi o que considerou um algoritmo AdaBoost.
- O algoritmo GradientBoost foi considerado para 40 e para 20 iterações, nas tabelas apresentadas, o resultado reflete 20 iterações sendo que não houve uma melhoria significativa entre ambas.
- A validação cruzada foi mais alta para o primeiro conjunto de variáveis mostrado.
- A certo ponto foi criada a variável DifHands sendo que quando ambos jogadores utilizam mãos opostas têm uma maior probabilidade de jogarem mais sets.
- Ao olhar para as curvas ROC, pode-se ver que os modelos num modo geral prevê ligeiramente melhor do que um classificador aleatório (com 50% de probabilidade de acerto).