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;

Performance dos modelos com o conjunto de variáveis nº 1

Variáveis usadas
DifNumberWins
DifRank
DifAge
DifHeight
Ground-clay
Ground-hard
Prize

Modelos	Accuracy	Precision	Recall	F1-score	Support	Mean ROC
Modelo 1- XGBoost	0.61	0.66 0.32	0.83 0.16	0.73 0.21	1645 857	0.53
Modelo 2- DecisionTree	0.54	0.65 0.35	0.65 0.35	0.65 0.35	1625 877	-
Modelo 3 - GradientBoosting	0.65	0.65 0.00	1.00 0.00	0.79 0.00	1625 877	0.53
Modelo 4- AdaBoost	0.65	0.65 0.54	1.00 0.01	0.79 0.02	1625 877	0.54

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 no entanto em termos de sensibilidade de previsão de sets de 3, o modelo 2 é o melhor;
- O modelo 3 não prevê sets de 3.

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	0.61	0.65 0.38	0.85 0.17	0.74 0.23	1625 877	0.51
Modelo 2 - DecisionTree	0.55	0.67 0.35	0.65 0.37	0.66 0.36	1652 850	-
Modelo 3 - GradientBoosting	0.66	0.66 0.00	1.00 0.00	0.80 0.00	1652 850	0.54
Modelo 4- AdaBoost	0.66	0.66 0.17	1.00 0.00	0.79 0.00	1652 850	0.53

Cross Validation com Random forest:
- 5 fold: 0.606
- 10 fold: 0.609

Observações:
- O melhor modelo é o 1, com uma performance mais alta do que o modelo anterior;
- Os modelos 3 e 4 não prevê sets de 3.

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	0.62	0.66 0.35	0.85 0.16	0.75 0.22	1652 850	0.52
Modelo 2 - DecisionTree	0.56	0.67 0.36	0.65 0.39	0.66 0.37	1657 845	-
Modelo 3 - GradientBoosting	0.66	0.66 0.00	1.00 0.00	0.80 0.00	1657 845	0.55
Modelo 4- AdaBoost	0.66	0.66 0.45	0.99 0.02	0.79 0.04	1657 845	0.54

Cross Validation com Random forest:

- 5 fold: 0.618
- 10 fold: 0.628

Observações:

- O melhor modelo é o 1;
- O modelo 3 não prevê sets de 3.
- O modelo 4 tem uma sensibilidade muito baixa a sets de 3.

Performance dos modelos com o conjunto de variáveis nº 4

Variáveis usadas
DifNumberWins
DifRank
DifAge
DifHeight
Prize
DifHands

Modelos	Accuracy	Precision	Recall	F1-score	Support	Mean ROC
Modelo 1- DecisionTree	0.57	0.68 0.37	0.66 0.39	0.67 0.38	1667 835	-
Modelo 2 - GradientBoosting	0.66	0.67 0.37	0.99 0.01	0.80 0.02	1667 835	0.53
Modelo 3- AdaBoost	0.66	0.67 0.37	0.98 0.02	0.79 0.04	1667 835	0.52
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:
- O modelo 2 e 3 têm uma sensibilidade muito baixa a sets de 3.

Performance dos modelos com o conjunto de variáveis nº 5

Variáveis usadas
DifNumberWins
DifRank
DifAge
DifHeight
Prize
DifHands
Ground_Clay
Ground_Hard

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

Cross Validation com Random forest:
- 5 fold: 0.621
- 10 fold: 0.616

Observações:
- O modelo 4 é o melhor modelo, sendo que tem uma boa sensibilidade a sets de 2 e 3 e tem uma accuracy mais alta do que o modelo 1.

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 4 é o melhor modelo, sendo que tem uma boa sensibilidade a sets de 2 e 3 e tem uma accuracy mais alta do que o modelo 1.
- Os modelos 2 e 4 têm uma sensibilidade muito baixa a sets de 3.

Considerações finais

- Os melhores modelos utilizam o algoritmo XGBoost e DecisionTree;
- Apesar de maioria dos modelos ter uma accuracy que vai de 50% a 70%, a melhor accuracy conseguida foi 62%;
- Há um tradeoff entre sensibilidade na previsão de sets de 3 e de accuracy em relação ao algoritmo DecisionTree e XGBoost.
- 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).