



Modelling: parte 2

Ténis na Bélgica

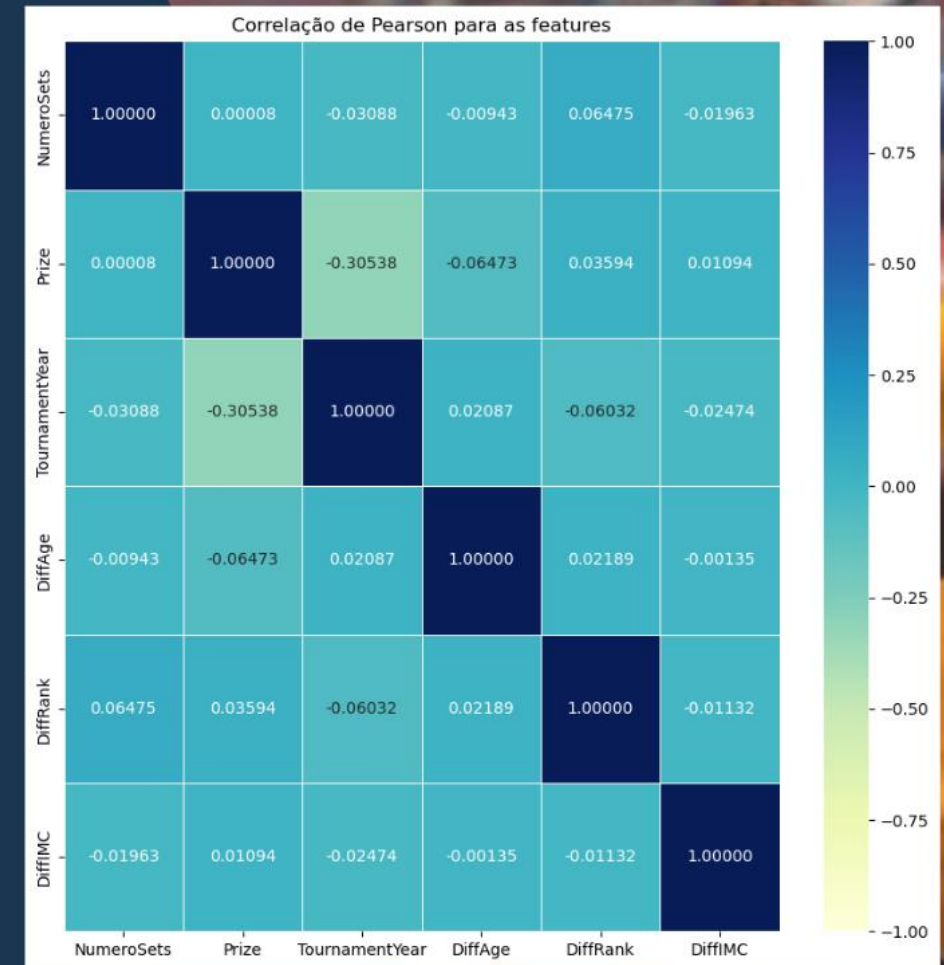
Grupo 4:

ALLAN KARDEC, N° 103380, TURMA CDB1
DIOGO FREITAS, N°104841, TURMA CDB1
JOÃO BOTAS, N°104782, TURMA CDB1
RICARDO ÂNGELO, N° 104826, TURMA CDB1

Correlações e Eta para novos modelos

- Adicionado o Rank aos modelos
- Eta para as variáveis qualitativas **vs** o target
- Alteração da matriz de correlação com as variáveis de player e oponente conjuntas

```
Coeficiente de Eta entre o NumeroSets e GroundNumerico: 0.0001557103212784806  
Coeficiente de Eta entre o NumeroSets e QTDBelgas: 0.0004835429299652118  
Coeficiente de Eta entre o NumeroSets e Finals: 0.000856455054545519  
Coeficiente de Eta entre o NumeroSets e SumMainHand: 0.00041773447172136884  
Coeficiente de Eta entre o NumeroSets e SumTrainer: 0.00040851156758612065
```



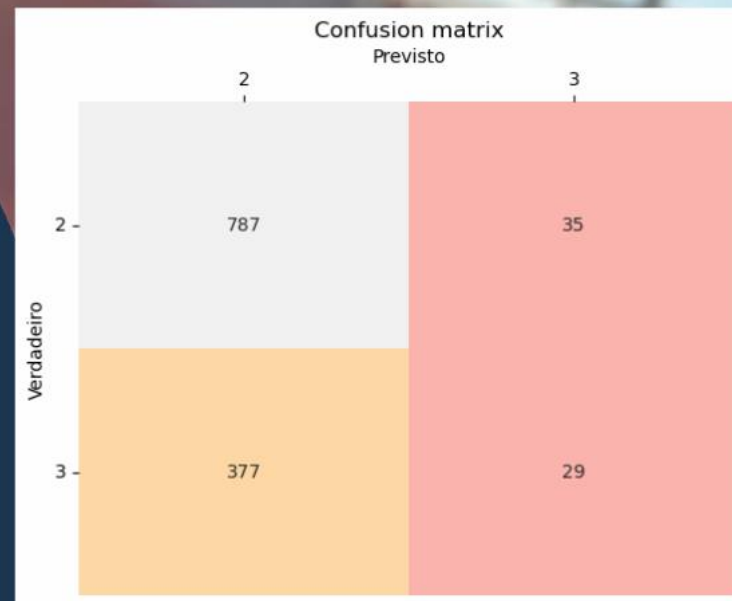
Outros modelos

- Logistic Regression e Knn com cross validation, média de 0.680 e de 0.622 de accuracy, respetivamente
- Modelo de Gradient Boosting com hiperparâmetros (através de escolha nas listas):
 - taxa de aprendizagem: 0.005
 - subsample: 0.7
 - número de árvores: 1250
 - profundidade e features máximas: 6
 - min_size e min_cut (nó folha): 6 e 7

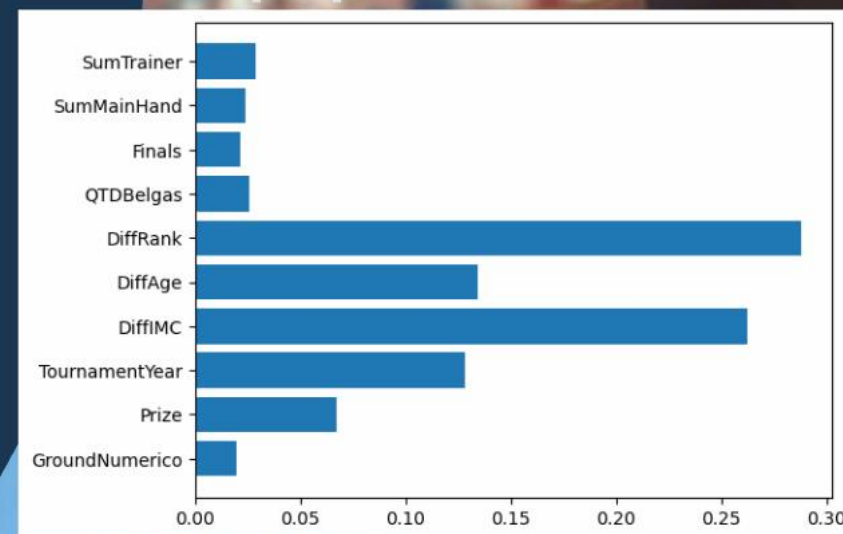
Métricas de classificação



Accuracy: 0.664
Recall: 0.957
Precision: 0.676
AUC: 0.564



Modelo de Gradient Boosting com hiperparâmetros

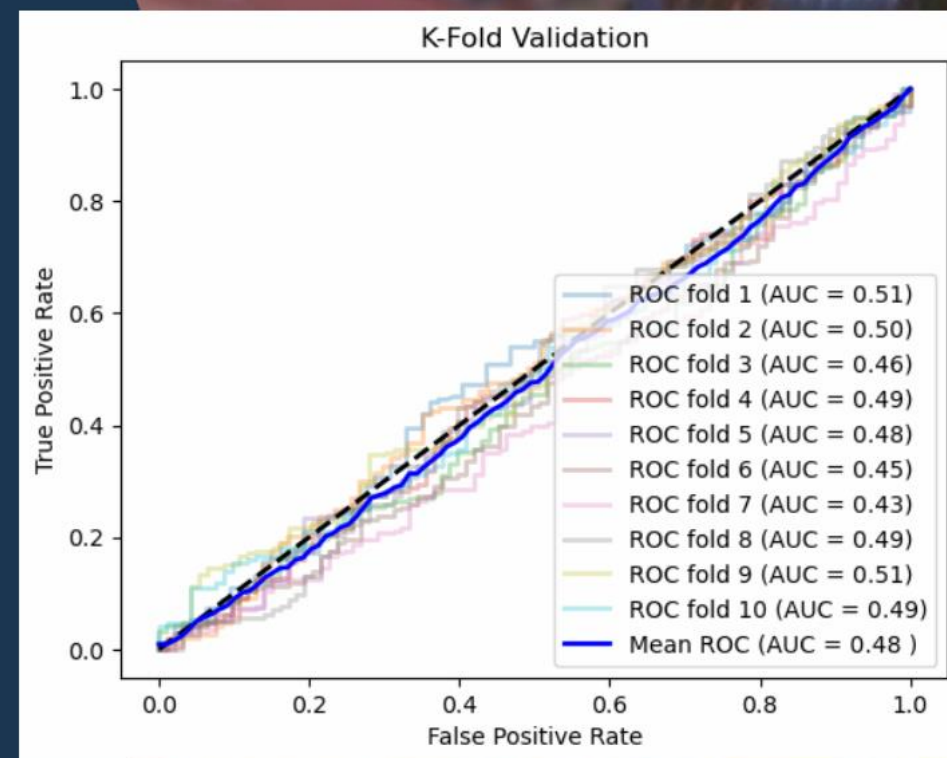


Modelo com estratificação K-fold

- Utilizada a função KFold da biblioteca sklearn, com o número de cortes = 10
- Modelo de Gradient Boosting
- AUC < 0.5 mas que são próximos a 0.5 (inverso)

Resultados pela média das métricas nas partições:

- Mean Accuracy: 0.649
- Mean Precision: 0.454
- Mean Recall: 0.493
- Mean F1-score: 0.421



Sets melhor de 5

- Divisão 70 treino / 30 teste
- Poucas observações no conjunto de teste (45 apenas)

Resultados Random Forest:

- Accuracy: 0.578
- Recall: 0.397
- Precision: 0.399
- AUC: 0.670

Resultados Gradient Boosting:

- Accuracy: 0.511
 - Recall: 0.342
 - Precision: 0.339
 - AUC: 0.596
-
- Modelo com poucas observações, difícil de fazer previsão
 - Não será mais aprofundado com 5 sets, mas incluído no relatório

		Confusion matrix		
		Previsto		
		3	4	5
Verdadeiro	3	23	3	5
	4	8	2	0
	5	2	1	1

		Confusion matrix		
		Previsto		
		3	4	5
Verdadeiro	3	21	7	3
	4	9	1	0
	5	2	1	1

Conclusão:

- Atualização das features e inclusão do rank nos modelos
- Modelos novos com hiperparâmetros e outras técnicas de partição
- Jogos à melhor de 5 sets

Para a próxima semana...

- Talvez testar mais um ou outro modelo
- Mais gráficos com as features selecionadas
- Tirar conclusões dos resultados obtidos

- Nota:

Existem, aproximadamente, 800 jogadores que possuem os seus respetivos nomes escritos de forma diferente no site do ATP, no site do Tennis Explorer e na base de dados inicial

2000 Alexandr Dolgoplov

NaN

NaN

Right-Handed

Dolgoplov Aleksandr - profile



Dolgoplov Aleksandr

Country: Ukraine
Height / Weight: 180 cm / 71 kg
Age: 34 (7. 11. 1988)
Current/Highest rank - singles: - / 13.
Current/Highest rank - doubles: - / 42.
Sex: man
Plays: right