APRENDIZADO DE MÁQUINA: TRABALHO FINAL 1

LYNCOLN SOUSA

OBJETIVO

- O objetivo do trabalho é utilizar técnicas de aprendizado de máquina na base de dados fornecida sobre NBA para classificar possíveis promissores atletas.

VARIÁVEIS

Var_name	Description
Name	Name
GP	Games Played
MIN	Minutes Played
PTS	PointsPerGame
FGM	FieldGoalsMade
FGA	FieldGoalsAttemps
FG%	FieldGoalPercent
3P MADE	3PointMade
3PA	3PointAttemps
3P%	3PointesPercent
FTM	FreeThrowMade
FTA	FreeThrowAttempts
FT%	FreeThrowPercent
OREB	OffensiveRebounds
DREB	DefensiveRebounds
REB	Rebounds
AST	Assists
SLT	Steals
BLK	Blocks
TOV	Turnovers
TARGET_5Yrs	Outcome: 1 if carrer length > 5yrs, 0 else

- Em geral, a base de dados possui informações sobre jovens atletas que conseguiram chegar, ou não, a 5 anos jogados na NBA. O empresário afirma que se o indivíduo jogou 5 anos na NBA, ele é talentoso.
- A base de dados fornecida apresenta 1350 observações e 21 variáveis, 19 são quantitativas enquanto 2 são qualitativas.

METODOLOGIA

- Foi gerado algumas medidas descritivas da base de dados, tais como: minimo, máximo, medidas de posição e quantidade de dados faltantes por variáveis.
- O modelo GBM (Gradient Boosting Machine) foi utilizado para realizar as classificações. Esse é um modelo que utiliza árvores de decisão para criação de um classificador ou regressor.

PRÉ PROCESSAMENTO

- Foi observado existência de dados faltantes, para isso foi utilizado o método de imputação KNN com k=5, assim, cada dado faltante será preenchido com a média de seus 5 vizinhos mais próximos.
- Também foi verificado a existência de variáveis com variância 0 ou quase 0 (nzv), pois variáveis assim não acrescentam informação para modelagem. Nenhuma das variáveis analisadas foram classificadas como nzv.

- Foram realizados cálculos para determinar as correlações entre as variáveis, as que obtiveram maior que 85% foram descartadas pois podem prejudicar a modelagem. Dentre todas as 21, foram excluídas 6, elas são: PTS, MIN, FGM, FTM, REB, X3P_MADE.
- Também foi verificado a existência de dependência linear entre as variáveis, isso é, se uma variável é combinação linear de uma outra. Não foi encontrado dependência linear.

VALIDAÇÃO CRUZADO

- Com os dados tratados, foi realizado a modelagem de GBM com os seguinte hiperparâmetro:
- Profundidade = 1,2,3,5,8
- Número de árvores = 500,1000,1500,2500
- Coeficiente de aprendizado = 0.01,0.02,0.03

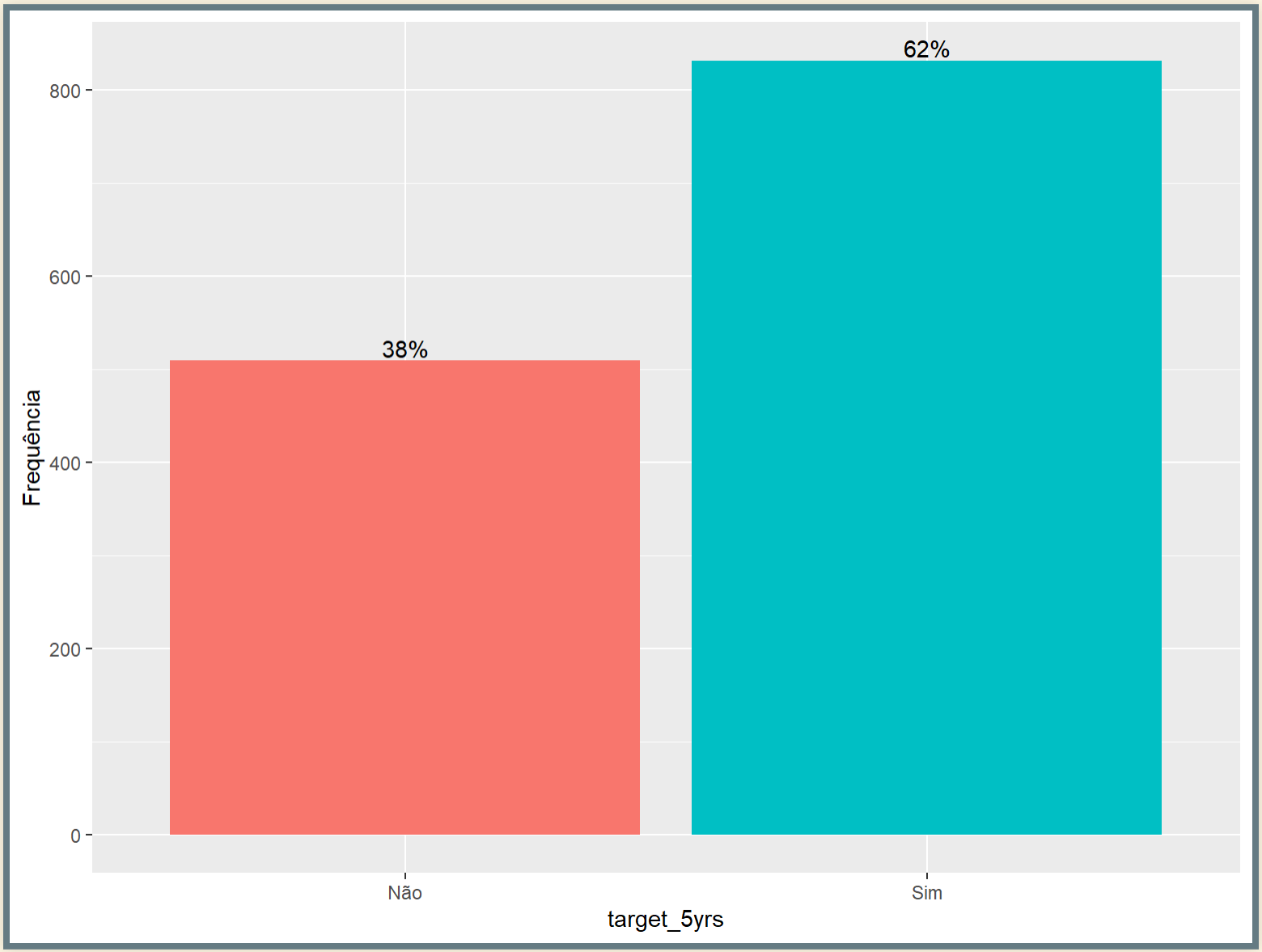
- Para a realização de reamostragem, foram escolhidos os métodos de bootstrap com 10 reamotragens e kfold repetido com 10 reamostragens e 3 repetições. O objetivo é verificar qual é o melhor método.
- O desempenho de cada modelo será analisado baseado no resultado do seu AUC.
- Dentre os 120 modelos testados, aquele com melhor desempenho será escolhido para a realização da modelagem final.

RESULTADOS

TABELA DE MEDIDAS DESCRITIVAS

	gp	min	pts	fgm	fga	fg	x3p_made	x3pa	x3p	ftm	fta	ft	oreb	dreb
Min.	11.0	3.1	0.70	0.30	0.80	23.8	0.000	0.000	0.0	0.0	0.00	0.0	0.00	0.20
1st Qu.	47.0	10.9	3.70	1.40	3.30	40.2	0.000	0.000	0.0	0.6	0.90	64.7	0.40	1.00
Median	63.0	16.1	5.55	2.10	4.80	44.1	0.100	0.300	22.4	1.0	1.50	71.2	0.80	1.70
Mean	60.4	17.6	6.80	2.63	5.89	44.2	0.248	0.779	19.3	1.3	1.82	70.3	1.01	2.03
3rd Qu.	77.0	22.9	8.80	3.40	7.50	47.9	0.400	1.200	32.5	1.6	2.30	77.6	1.40	2.60
Max.	82.0	40.9	28.20	10.20	19.80	73.7	2.300	6.500	100.0	7.7	10.20	100.0	5.30	9.60
NA	0.0	0.0	0.00	0.00	0.00	0.0	0.000	0.000	11.0	0.0	0.00	0.0	0.00	0.00

PROPORÇÃO DE CLASSIFICAÇÃO



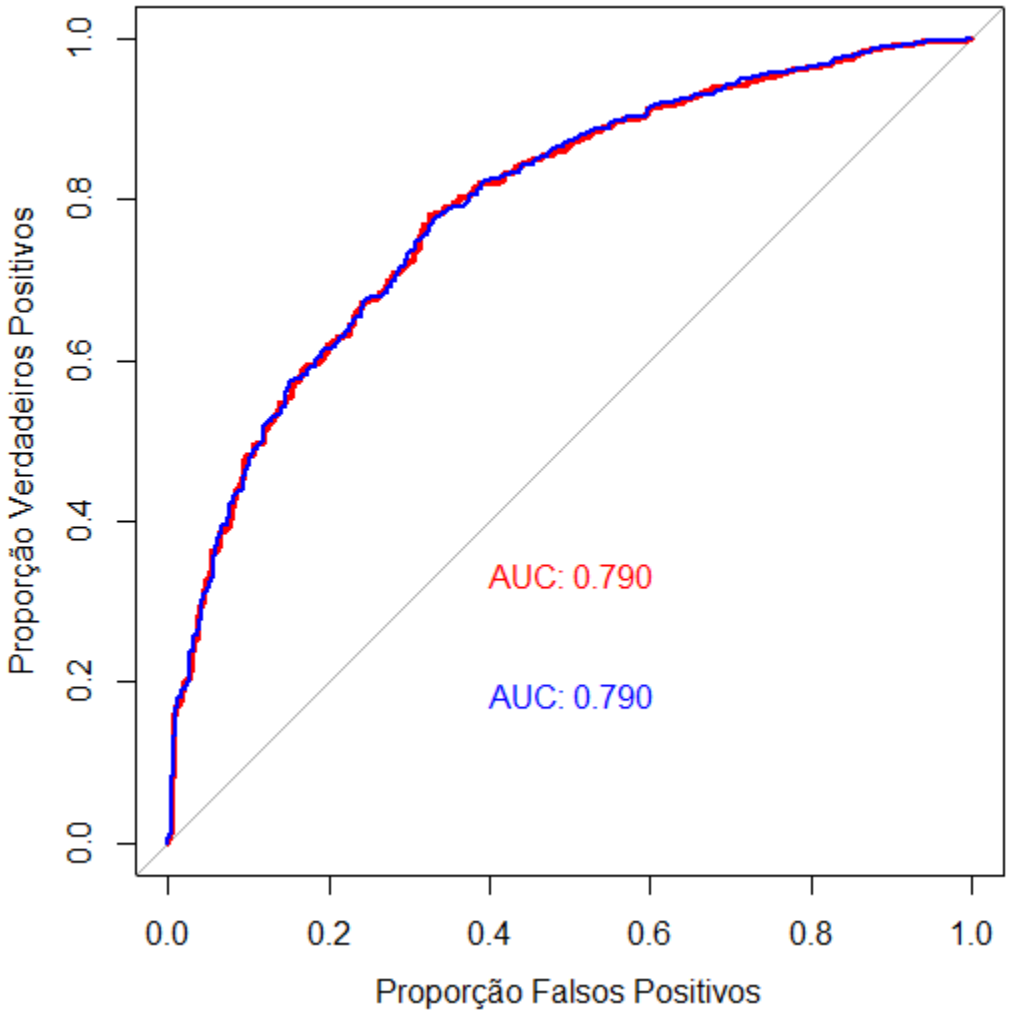
MODELAGEM POR BOOTSTRAP

shrinkage	interaction.depth	n.minobsinnode	n.trees	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
0.02	1	10	500	0.7360742	0.4897487	0.8083469	0.0179425	0.0582529	0.0309960
0.01	1	10	1000	0.7356544	0.4928804	0.8073086	0.0182685	0.0564141	0.0300270
0.01	1	10	500	0.7355423	0.4700280	0.8139597	0.0190183	0.0573608	0.0339551
0.03	1	10	500	0.7349641	0.5011345	0.8016327	0.0180686	0.0541201	0.0303905
0.01	1	10	1500	0.7344759	0.4969571	0.7979237	0.0185185	0.0586054	0.0331156
0.01	2	10	500	0.7331509	0.5005241	0.7976356	0.0177648	0.0453508	0.0322626
0.02	1	10	1000	0.7330140	0.5029127	0.7979219	0.0189153	0.0556718	0.0321356
0.01	3	10	500	0.7311117	0.5082253	0.7926226	0.0183844	0.0439509	0.0316149
0.01	1	10	2500	0.7301538	0.4999779	0.7926617	0.0191687	0.0605348	0.0301202
0.01	2	10	1000	0.7299312	0.5083512	0.7917321	0.0188368	0.0448367	0.0321288
0.02	2	10	500	0.7293964	0.5122322	0.7894025	0.0186412	0.0433278	0.0329799
0.02	1	10	1500	0.7279637	0.4950320	0.7937183	0.0198608	0.0496167	0.0260282
0.03	1	10	1000	0.7277687	0.4992030	0.7963155	0.0197205	0.0516384	0.0274936
0.01	5	10	500	0.7272565	0.5147278	0.7859639	0.0179933	0.0474704	0.0259954

MODELAGEM POR REPEATEDCV

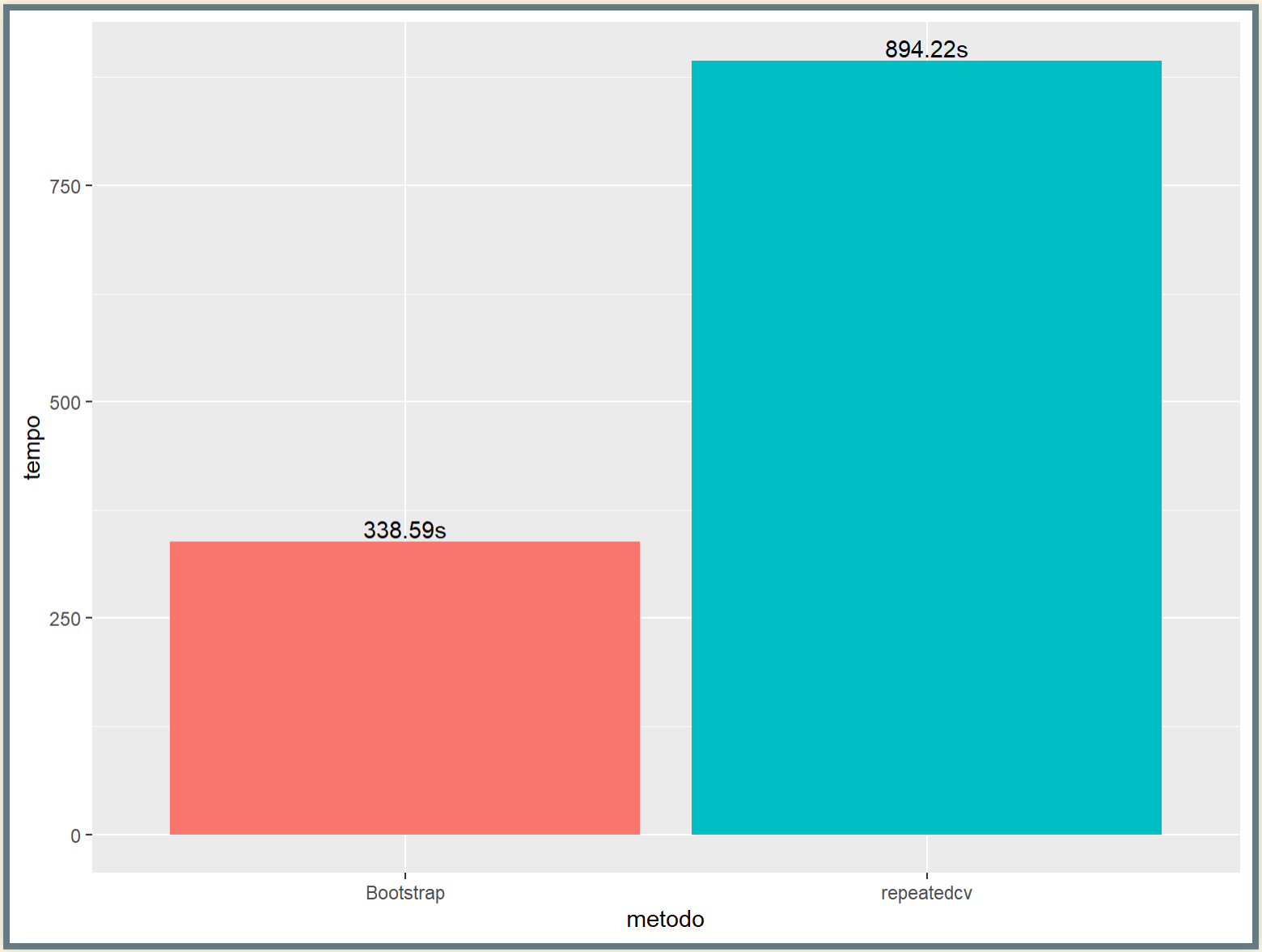
shrinkage	interaction.depth	n.minobsinnode	n.trees	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
0.01	1	10	1000	0.7508213	0.5101961	0.8187177	0.0362342	0.0766964	0.0525136
0.02	1	10	500	0.7508041	0.5108627	0.8175081	0.0364745	0.0738667	0.0526466
0.01	1	10	1500	0.7504213	0.5141961	0.8159113	0.0354565	0.0766966	0.0533919
0.03	1	10	500	0.7500147	0.5076209	0.8135064	0.0353040	0.0756852	0.0518031
0.02	1	10	1000	0.7487773	0.5181176	0.8135016	0.0355222	0.0813163	0.0486657
0.01	1	10	500	0.7486439	0.4996863	0.8263291	0.0365873	0.0740602	0.0504586
0.01	2	10	500	0.7478479	0.5121699	0.8155144	0.0363304	0.0786565	0.0497134
0.01	1	10	2500	0.7478209	0.5181307	0.8110968	0.0356295	0.0836920	0.0487916
0.02	1	10	1500	0.7463312	0.5168105	0.8078839	0.0350520	0.0843037	0.0474921
0.03	1	10	1000	0.7462945	0.5206928	0.8082760	0.0344498	0.0793720	0.0496582
0.01	2	10	1000	0.7462113	0.5174641	0.8094712	0.0356028	0.0764542	0.0505733
0.02	2	10	500	0.7459704	0.5154902	0.8078696	0.0359644	0.0779885	0.0533902
0.01	3	10	500	0.7459406	0.5135033	0.8106760	0.0368647	0.0800300	0.0481546
0.01	5	10	500	0.7434201	0.5167712	0.8002486	0.0370602	0.0774463	0.0532303

ROC DOS 2 MELHORES MODELOS



Azul: Bootstrap (0.7902) Vermelho: repeatedcv (0.7904)

TEMPO DE PROCESSAMENTO



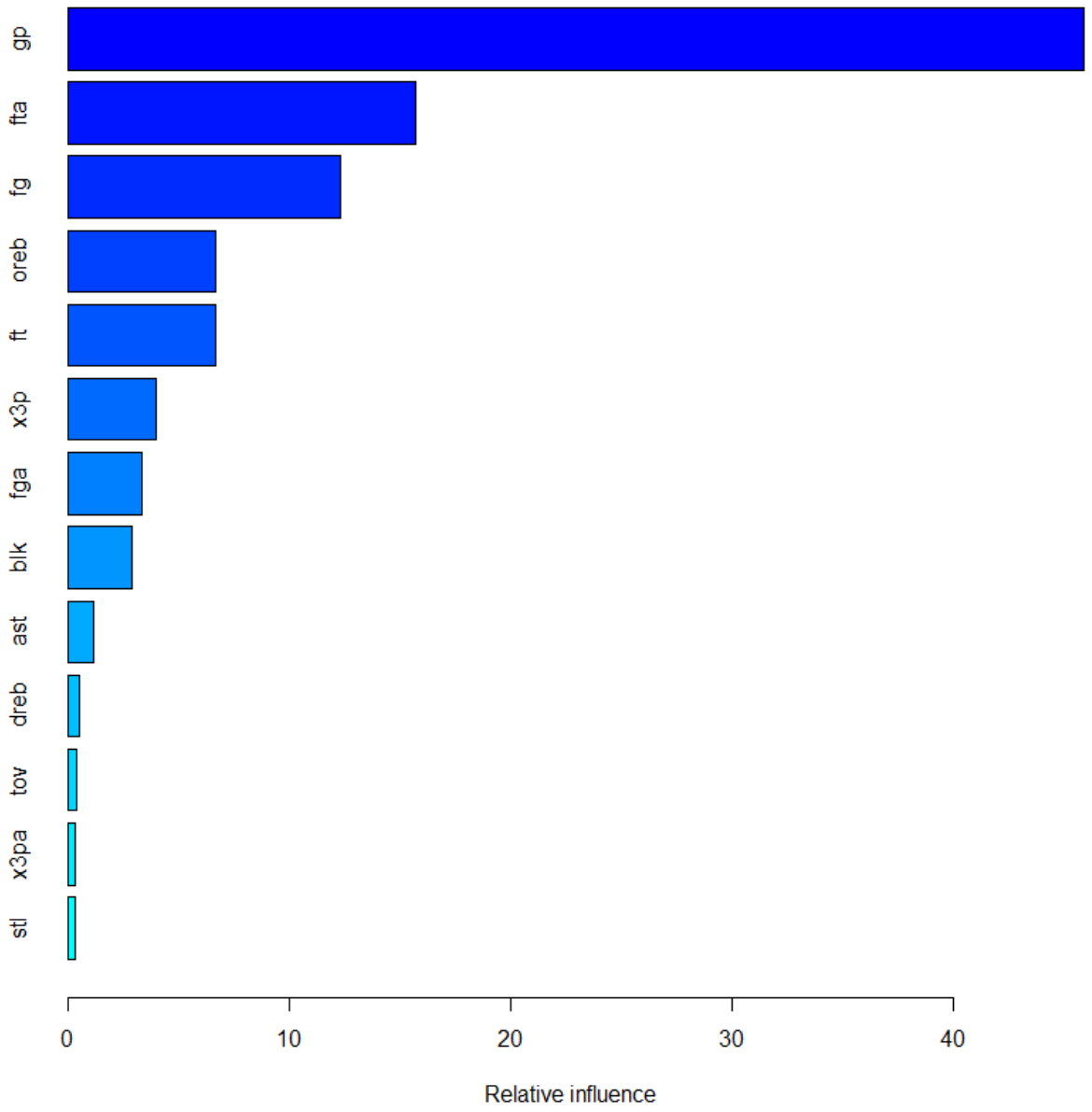
MODELO ESCOLHIDO

- O modelo que teve a melhor performance foi o de coeficiente de aprendizado 0.01, profundidade 1 e número de árvores 1000, com reamostragem por kfold repetido.
- Foi realizada a separação conforme a exigência de 25% para amostra teste e 75% para amostra treino. A modelagem a partir da amostra treino obteve o seguinte resultado:

interaction.depth	n.trees	shrinkage	n.minobsinnode	ROC	Sens	Spec	ROCSD	SensSD	SpecSD
1	1000	0.01	10	0.7615706	0.5305443	0.8177931	0.044361	0.072431	0.0497537

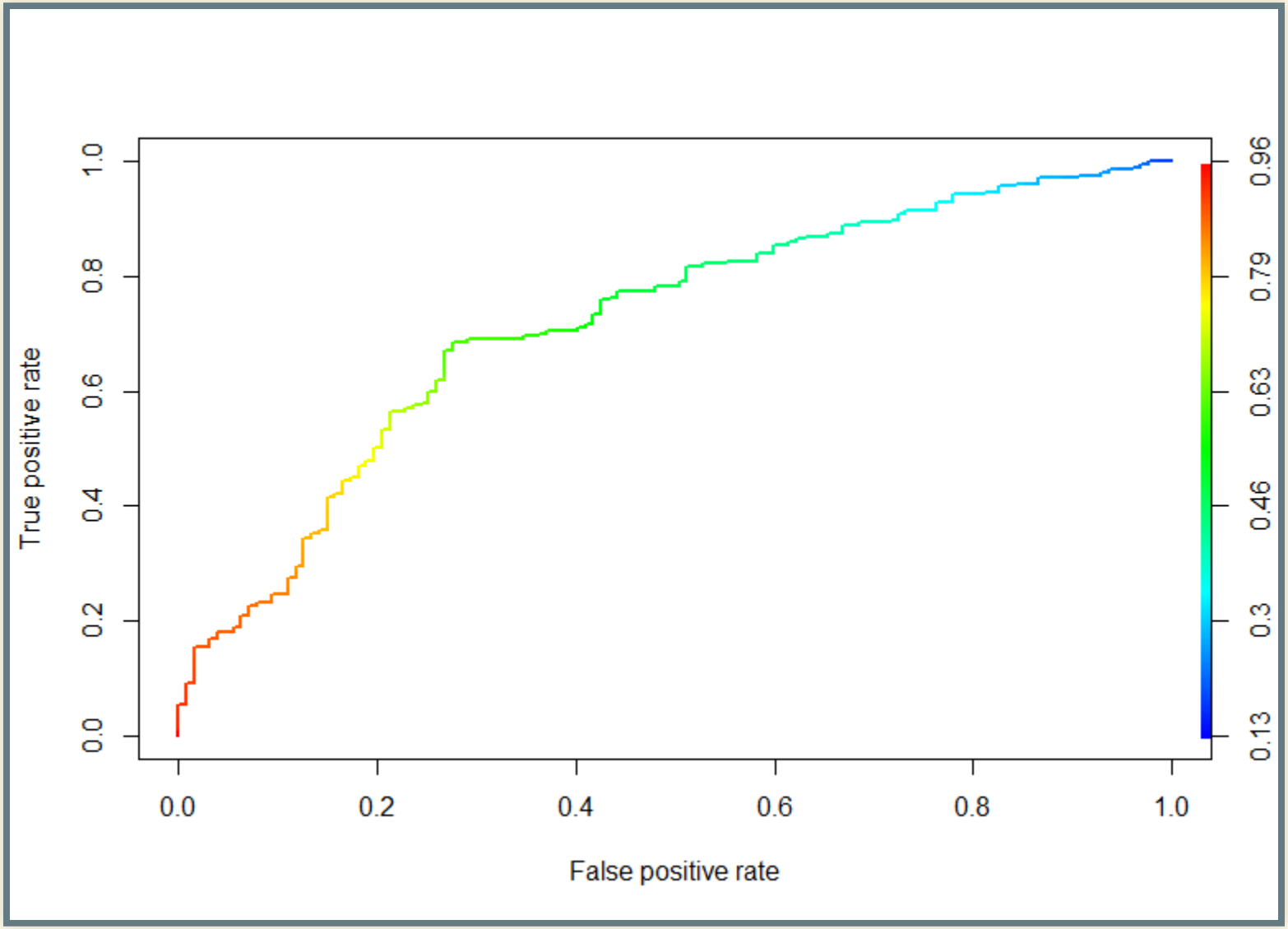
- Variáveis mais influentes

Var	Influence
gp	45.9172748
fta	15.6987368
fg	12.2980805
oreb	6.6680500
ft	6.6427327
x3p	3.9906534
fga	3.3055565
blk	2.8876777
ast	1.1636744
dreb	0.4858196
tov	0.3420671
x3pa	0.3033329
stl	0.2963437


gp
fta
fg
oreb
ft
x3p
fga
blk
ast
dreb
tov
x3pa
stl
0
10
20
30
40
Relative influence

- Aplicando a modelagem na amostra teste, é adquirido um vetor de probabilidades, então assim é necessário adotar um ponto de corte para a classificação.
- Foi utilizado uma função otimizadora que a partir dos dados da predição retorna o melhor ponto de corte, que no caso foi de 0.6033771.

CURVA ROC


True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
0.13
0.3
0.46
0.63
0.79
0.96

RESULTADO PARA AMOSTRA TESTE

- Aplicando a modelagem na amostra teste e definindo o ponto de corte, são obtidos os seguintes resultados:

Acurácia	Sensibilidade	Especificidade	AUC
0.7006	0.7244	0.686	0.7178

CONCLUSÃO

- Das modelagens feitas, as que mais se destacaram foram aquelas realizadas por kfold repetido.
- Apesar do método kfold ser mais demorado, ele foi escolhido por possuir maiores AUCS em várias modelagens comparados aos de bootstrap.
- O modelo final apresentou Acurácia em torno de 70%, isso significa que o modelo possui uma boa taxa de acerto.
- O modelo apresentou Sensibilidade um pouco maior que a Especificidade, isso mostra que ele acerta mais os verdadeiros positivos do que os verdadeiros negativos.