Projeto Final | Aprendizado de Máquina

LEONARDO DE OLIVEIRA PENNA JEFERSON PEREIRA DE ANDRADE
NATHALIA STEFANY SANTOS SILVA MARIANA DE CASTRO PASQUALINI
MIGUEL GIOVANE GONZAGA RODRIGUES

09 de dezembro de 2022

1. Classificador

Problema

Temos um problema de classificação multiclasse. Como cada jogador tem mais de uma classificação considerada correta, uma abordagem possível é utilizando modelos *multilabel*, que pode atribuir um ou mais rótulos não-exclusivos para uma mesma observação.

Porém, para simplificar o problema, optamos por tratá-lo apenas como um classificador multiclasse, que atribui dentre as N classes disponíveis, um único rótulo. Para não perder a informação de que mais de uma posição pode estar correta, treinamos o modelo no formato longo e para a predição final, agrupamos por jogador a de maior probabilidade.

Y é a variável resposta categórica, com $|\mathcal{C}| = 27$, em que estimamos $\mathbb{P}(Y = c|\mathbf{x})$. Temos uma matriz de covariáveis X com 41 features.

Função de risco

$$R(g) := \mathbb{E}[\mathbb{I}(Y \neq g(X))] = \mathbb{P}(Y \neq g(X))$$

Adotamos a função de perda 0-1, comum para problemas de classificação.

Data splitting

O conjunto de dados de treino foi separado em treino e validação, para estimar $\hat{R}(g)$. 70% das observações foram para treino e 30% para validação.

Análise descritiva

Na tabela abaixo, encontram-se as principais estatísticas descritivas das covariáveis dos modelos.

Tipo	Variável	Média	Desvio-padrão	Mínimo	Q25	Q50	Q75	Máximo
numeric	rating	69.7	7.6	45	65	70.0	75	94
numeric	height	182.0	6.5	162	178	182.0	187	207
numeric	weight	75.7	6.9	48	71	75.0	80	110
numeric	age	25.0	4.6	17	21	25.0	28	40

Tipo	Variável	Média	Desvio-padrão	Mínimo	Q25	Q50	Q75	Máximo
numeric	weak_foot	3.0	0.7	1	3	3.0	3	5
numeric	skill_moves	2.4	0.8	1	2	2.0	3	5
numeric	ball_control	61.9	17.8	9	57	67.0	74	95
numeric	dribbling	58.4	20.0	4	51	64.0	72	97
numeric	marking	46.2	23.2	4	22	51.0	67	92
numeric	sliding_tackle	47.9	23.2	6	24	55.0	68	95
numeric	standing_tackle	50.1	23.4	8	27	57.0	71	92
numeric	aggression	58.4	18.0	11	46	62.0	72	94
numeric	reactions	65.5	9.9	30	58	66.0	73	96
numeric	attacking_position	52.3	20.8	3	40	58.0	68	94
numeric	interceptions	49.4	22.1	5	27	54.0	69	93
numeric	vision	56.0	15.4	10	45	58.0	68	94
numeric	composure	59.4	14.3	5	50	62.0	70	94
numeric	crossing	53.0	19.6	7	41	58.0	68	90
numeric	short_pass	61.9	15.6	11	56	66.0	73	92
numeric	long_pass	55.9	16.1	10	46	59.5	68	93
numeric	acceleration	66.3	14.4	15	59	68.0	77	96
numeric	speed	66.5	14.2	11	59	69.0	76	96
numeric	stamina	64.6	15.8	12	58	68.0	75	95
numeric	strength	66.3	12.5	25	58	68.0	75	94
numeric	balance	64.2	13.9	12	56	66.0	74	95
numeric	agility	64.6	14.5	16	56	67.0	75	96
numeric	jumping	66.2	11.2	25	59	67.0	74	95
numeric	heading	54.7	18.6	5	47	58.0	68	93
numeric	shot_power	59.3	18.2	3	49	64.0	73	93
numeric	finishing	47.5	20.6	2	30	51.0	65	95
numeric	long_shots	50.4	20.2	5	35	55.0	66	91
numeric	curve	50.7	19.8	6	37	53.0	67	91
numeric	freekick_accuracy	45.9	18.7	7	32	46.0	61	91
numeric	penalties	51.3	16.5	10	41	53.0	64	92
numeric	volleys	46.3	19.0	5	32	48.0	61	93
numeric	gk_positioning	17.0	18.2	1	8	11.0	14	91
numeric	gk_diving	17.2	19.0	1	8	11.0	14	89
numeric	gk_kicking	16.7	17.4	1	8	11.0	14	95
numeric	gk_handling	16.9	18.2	1	8	11.0	14	91
numeric	gk_reflexes	17.3	19.3	1	8	11.0	14	90

Modelos

Foram treinados 5 modelos distintos:

- Regressão multinomial ou *softmax*
- Naive Bayes
- Support Vector Machines
- Árvores de decisão
- Random Forest

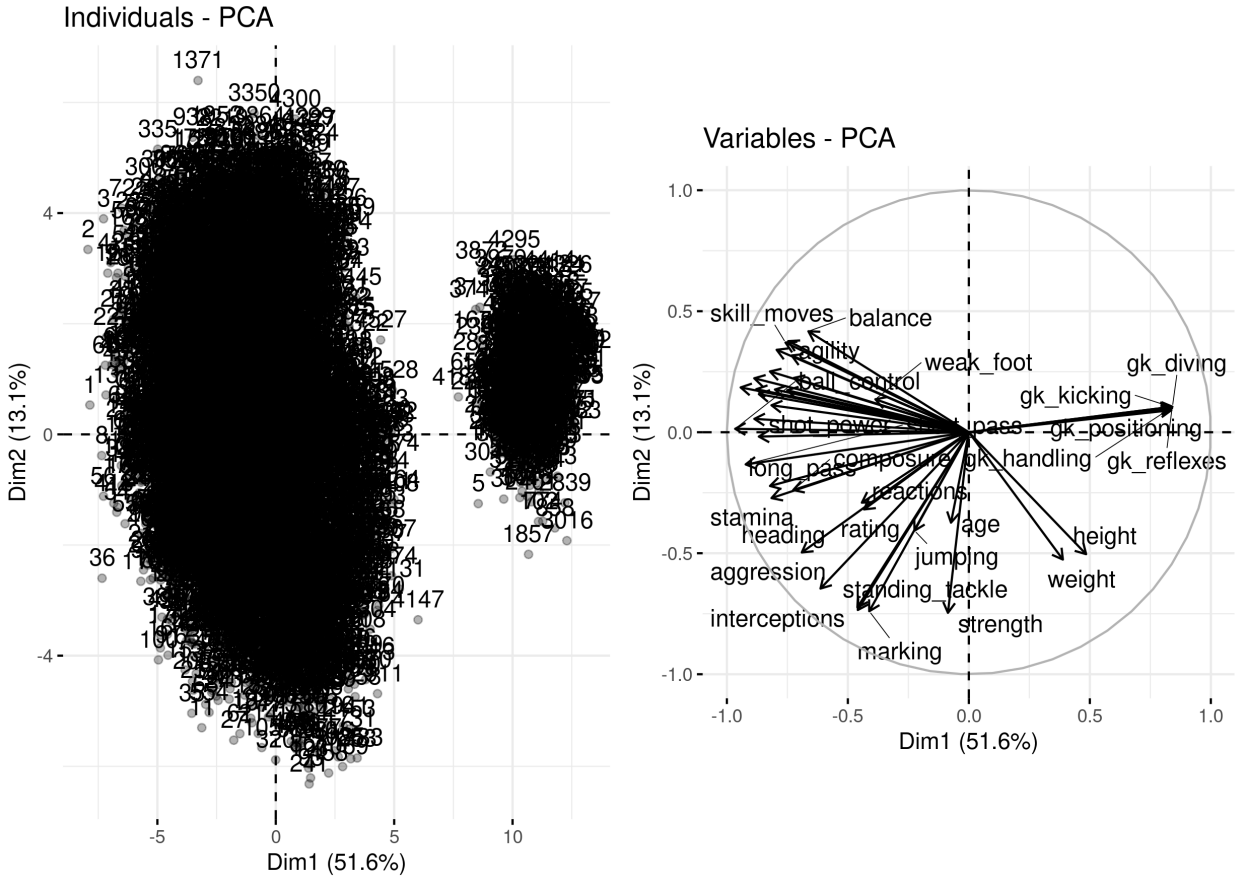
O método que forneceu o melhor resultado foi o Random Forest, de acordo com o risco estimado no conjunto de validação, apresentado na tabela a seguir.

Modelo	Risco estimado
Naive Bayes	0.3590325
SVM	0.1806500
Regressão multinomial	0.2781557
Árvore de decisão	0.3862434
Random Forest	0.0090703

É importante notar que, comparado aos outros métodos, é possível que o método de menor risco estimado esteja super-ajustando aos dados de validação.

Redução de dimensionalidade

Aplicamos a técnica de análise de componente principal (PCA) para redução de dimensionalidade das covariáveis numéricas e há claramente um agrupamento nos dados. Observa-se no gráfico abaixo os primeiros dois componentes, com as observações individuais e a correlação entre as variáveis.



Comentários

O ponto mais difícil foi fazer o tratamento da variável resposta `Preffered_Position`, quebrando as uma ou mais posições de cada jogador. Fazer uma seleção de variáveis pode ser interessante para melhorar o desempenho dos modelos. Com mais tempo, outro ponto é alterar os *tunning parameters* de cada modelo e também usar uma outra função de risco, como por exemplo a entropia cruzada para classificação multiclasse. Ainda, é possível explorar outras técnicas e modelos para atribuir mais de uma classe à mesma instância.

2. Recomendador