Estudo da influência de variáveis associadas a visibilidade visibilidade na evolução do número de adeptos

Questão: Qual o contributo, no número de praticantes por época, de variáveis relacionadas com a visibilidade da modalidade?

Considerações iniciais:

A visibilidade de uma determinada modalidade faz aumentar o interesse das pessoas na mesma e pode contribuir para um aumento do seu número de praticantes. Neste estudo pretende-se estudar a medida em que um conjunto de variáveis podem ajudar a explicar a evolução de números de praticantes ao longo de um conjunto de épocas. Estas variáveis encontram-se associadas à visibilidade da modalidade por gerarem notoriedade associada ao mérito desportivo ou à organização de determinados eventos.

A questão proposta refere o número de praticantes por época. A cada época este número é a soma de novas inscrições, continuações e desistência. Um possível, e interessante, caminho para esta análise teria sido tentar explicar a diferença de praticantes de época para época, ao invés do total por época. Optei por não o fazer por um conjunto de razões:

- Nenhuma desta variáveis determina directamente o número de praticantes ou variação de época para época, apenas os influenciam, em conjunto com muitas outras. Dado o pequeno número de variáveis e o reduzido número de observações torna-se muito difícil extrair a relação existente com a variação entre cada que é uma variável mais noisy do que o número total de praticantes.
- Apesar de as observações que compõem as diversas variáveis terem uma frequência de época não se pode determinar, e é certamente incorrecto, que a sua influência também o tenha. Por exemplo a organização de um evento como o Campeonato da Europa de 2003-2004 terá influenciado o número de

inscrições não apenas nessa época mas também em épocas anteriores e posteriores.

- As variáveis, e visibilidade, que influenciam o número de novas inscrições também influenciarão o número de permanências. Se em épocas com uma variação negativa é claro que o número de abandonos foi maior que o número de novas inscrições, em épocas com crescimento é impossível, com estes dados, perceber a contribuição relativa de novas inscrições e continuações, tornando os objectivos da análise menos claros.

1. Exploração das variáveis

Comecemos por olhar para a informação que temos em relação ao número de praticantes por época:

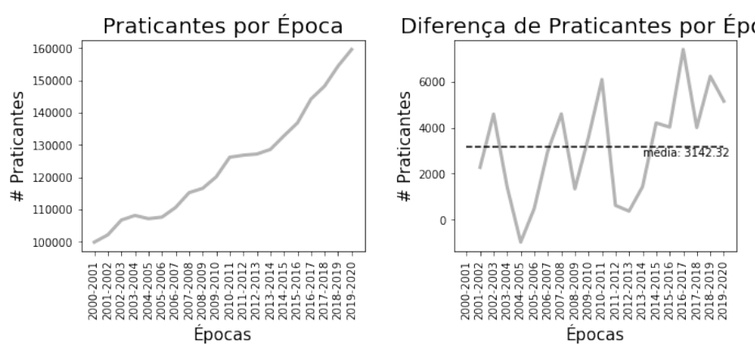

Fig1. Esquerda: evolução do número de praticantes por época. Direita: Diferença no número de praticantes de época para época.

O número de praticantes cresceu continuamente desde a época 2000-2001 à época 2019-2020, com excepção da época 2004-2005, um crescimento de praticamente 60% para o período. A média de crescimento foi de aproximadamente 3190 adeptos, por ano mas houve uma variação considerável no crescimento ao longo das épocas, desvio padrão de aprox 2359 praticantes. Algumas observações a destacar: 1 - desde época de 2014-2015 que o número de adeptos cresce acima da média, com um pico evidente em 2016-2017, a época seguinte à conquista por Portugal do Campeonato Europeu; 2 - houve dois períodos de notório crescimento mais lento, o primeiro entre as épocas 2003-2004 e 2005-2006, o segundo entre 2010-2011 e 2013-2014; 3 - a única época com crescimento negativo foi a de 2004-2005, curiosamente imediatamente a seguir ao Campeonato da Europa organizado por Portugal, o que, especulando, se poderá dever a uma regressão à média depois de vários anos de forte antecipação e investimento,

As variáveis que temos para tentar explicar a evolução do número de adeptos são:

- **Score :** o valor resultante de um scoring system que tem em consideração a posição alcançada em Campeonatos da Europa, Campeonatos do Mundo, bem como os prémios de melhor marcador e treinador.
- **Organizador:** se o país organizou alguma competição internacional nessa época.
- **Prémios:** prémios individuais recebidos pela seleção e seus jogadores.
- **Ranking:** ranking FIFA atingido pela seleção em cada uma das épocas.
- **IDH(** Índice de Desenvolvimento Humano**) :** medida comparativa usada para classificar os países pelo seu grau de desenvolvimento humano.

As quatro primeiras variáveis dizem directamente respeito ao futebol e à sua visibilidade, o IDH, no entanto, não. A variável poderá influenciar o número de praticantes, dado que na sua elaboração a métrica tem em conta o nível de educação e rendimento de um país, realidades que certamente poderão ter um efeito na prática desportiva no mesmo, mas isto não é necessariamente verdade e o IDH não está directamente relacionado com a visibilidade da modalidade, o foco principal da questão colocada. Continuarei a incluir a variável na análise, mas sempre que necessário realizarei análise separadas com e sem a mesma.

Como já mencionado, apesar de as observações corresponderem a uma determinada época estas têm uma influência que extravasará para épocas adjacentes. Com base nisto introduzi alterações em algumas das variáveis que passo a explicar:

- As observações nas variáveis *Score* e *Prémios* apenas podem influenciar as épocas seguintes àquelas em que foram obtidas. Para além disso a influência de um bom score ou prémio possui um certo momento que se estende, provavelmente, por mais do que uma época, interagindo com outras observações aí registadas. Seguindo este racional criei duas novas

variáveis *MScore* e e *MPrémios* que resultam de deslocar as variáveis originais uma época para o futuro seguido da aplicação de uma média rolante com uma janela de três épocas.

- A variável *Ranking* é ligeiramente diferente, uma vez que o ranking da FIFA vai mudando várias vezes ao longo do ano e, ao escolhermos um valor por época, perdemos informação sobre os outros valores da mesma época bem como sobre a sua evolução, tudo factores que podem influenciar de maneira diferente a visibilidade da modalidade. Decidi apenas aplicar uma média rolante de três épocas à variável, para amenizar um pouco os seus extremos e ser fiel a um espírito de continuidade.

- A organização de uma competição tem um efeito que abrange não apenas a época dessa competição mas também épocas anteriores e posteriores. Lembrar todo o esforço de preparação e entusiasmo que antecedeu a realização do Euro 2004, a festa que foi a época do evento e todas as novas infra-estruturas e entusiasmo que passaram para os anos seguintes. Para mimetizar esse efeito apliquei um filtro gaussiano, com uma janela de três épocas, e desvio padrão = 1, à variável. Desta forma o efeito da organização de uma competição estende-se às épocas anterior e posterior, com um valor mais baixo.

Olhemos então para as variáveis originais e criadas:

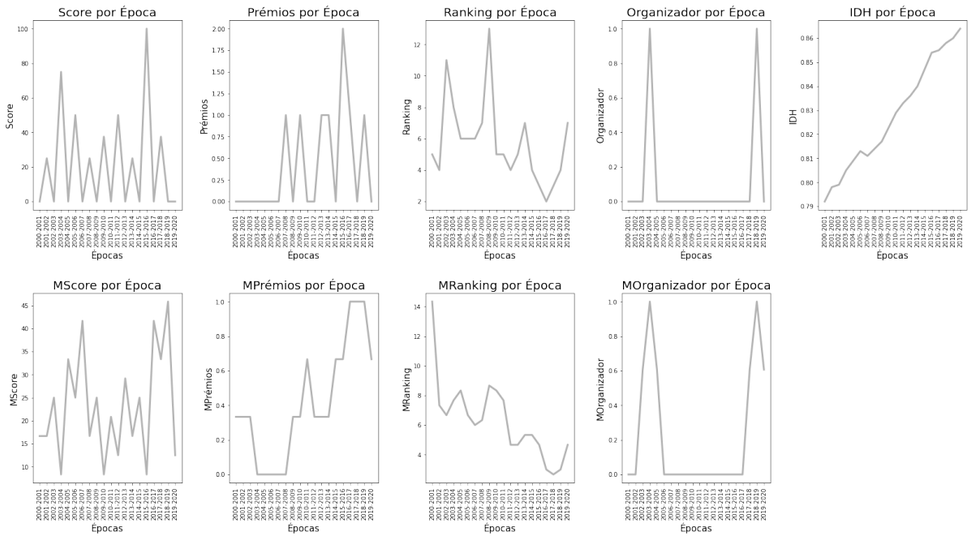

Fig2. Topo: variáveis originais. Fundo: Variáveis criadas.

É visível como as transformações operadas mudaram as variáveis originais tornando-as menos noisy e espalhando o efeito das suas observações por várias épocas, tornando a sua influência mais contínua. Várias apresentam agora padrões crescentes ou decrescentes, o que as pode tornar úteis na explicação da evolução do número de adeptos.

2. Relação entre as variáveis explicativas e o número de praticantes

Para determinar que variáveis podem ser mais úteis em explicar a evolução do número de praticantes ao longo das épocas pode-se começar por calcular a correlação entre a cada uma delas e o número de praticantes, .

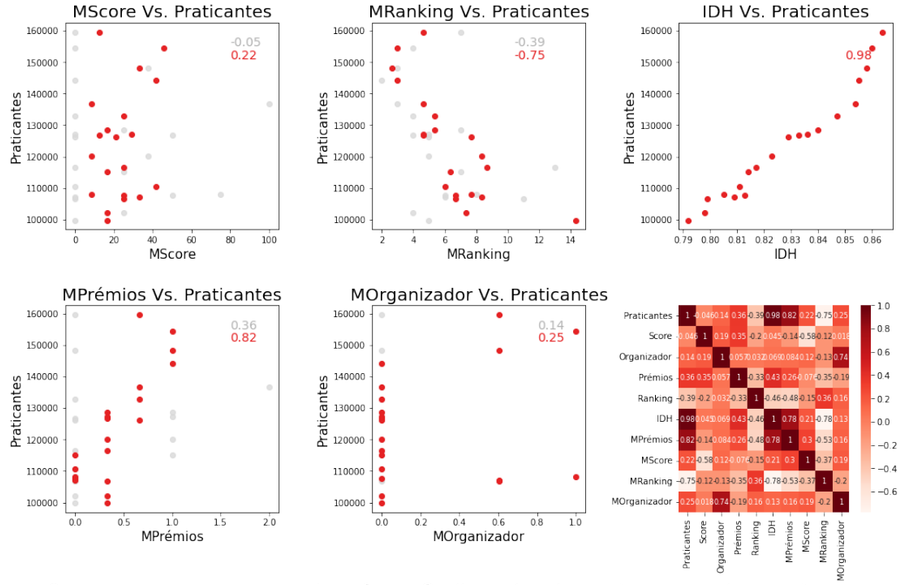

Fig3. Topo e primeiros dois subplots do fundo: Scatter plots entre variáveis explicativas e número de praticantes. Em cada a plot, tirando para *IDH*, podemos ver a cinzento a variável original e a vermelho a variável transformada. Os números representam os coeficientes de correlação para cada uma delas. Fundo à direita: matriz de correlações entre todas as variáveis.

Na figura acima pode-se constatar como as transformações introduzidas nas variáveis melhoraram consideravelmente a sua correlação com a variável alvo. *IDH*, *MRanking* e *MPrémios* apresentam correlações bastante fortes com o número de praticantes: 0.98, -0.75 e 0.82, respectivamente. As correlações com *MOrganizador* e *MScore* são mais fracas.

Para obter uma estimativa de quanta confiança se pode ter nestes valores realizei um bootstrapping, efectuando uma reamostragem com reposição, para gerar uma distribuição dos coeficientes de correlação entre cada uma das novas variáveis e o número de praticantes.

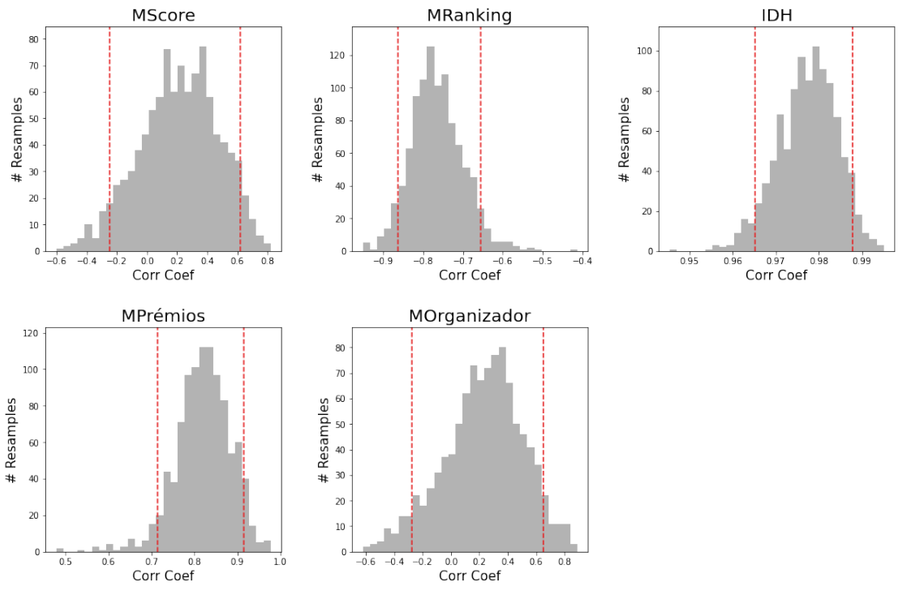

Fig4. Histogramas dos coeficientes de correlação obtidos para 1000 reamostragens, com reposição, de pares de pares de observações variável explicativa - variável alvo. A vermelho os limites inferiores e superiores para o intervalo de confiança que contém 90% dos valores da distribuição.

Pelo Teorema do Limite Central, um números grande de amostras independentes aleatórias tenderão a aproximar uma distribuição normal, mesmo que a distribuição da população de origem não o seja. Se tal fosse o caso, o valor expectável para o coeficiente de correlação entre cada uma das variáveis e o número de praticantes seria a média da distribuição e o centro do intervalo de confiança. As distribuições por nós obtidas são ligeiramente enviesadas, impossibilitando-nos de fazer essa afirmação de forma absolutamente clara. As distribuições de reamostragem geradas permitem-nos, no entanto, concluir que *IDH*, *MPrémios* e *Mranking* possuem associações fortes e de sinal claramente definido com a variável alvo, positivo os dois primeiros, negativo o segundo. O mesmo não acontece com *MScore* e *MOrganizador* já que para ambas variáveis o valor 0 se encontra contido no intervalo de confiança e a percentagem de reamostragens com um coeficiente de correlação menor ou igual a 0 é de 21 e 20% , respectivamente.

IDH, *MPrémios* e *MRanking* posicionam-se assim como bons candidatos para variáveis que podem ajudar a explicar a evolução do número de praticantes por época, sendo que, como já discutido, apenas as duas últimas se encontram directamente associadas à visibilidade da modalidade.

2. Efeito das variáveis explicativas no número de praticantes

A análise de correlações mostrou a força e o sinal da relação entre as variáveis explicativas e a variável alvo. Para perceber e quantificar o efeito que cada uma tem no número de praticantes recorri à regressão linear múltipla. Como é possível ver na matriz de correlações acima as variáveis que mais correlacionadas estão com a variável alvo, encontram-se também bastante correlacionadas entre si. É possível que a correlação de uma com o número de praticantes se deva à sua correlação com outra ou outras variáveis. A título especulativo, o Ranking FIFA de uma seleção está associado a bons desempenhos dessa seleção, bons desempenhos esses que resultarão numa maior probabilidade de obtenção de prémios. Mas qual variável terá um maior efeito na evolução do número de praticantes?

Para lidar com esta questão utilizarei uma regressão linear regularizada, em que a regularização vai encolher os coeficientes associados com cada variável em direção a 0, desta forma selecionando as variáveis que melhor explicam por si a variável alvo, gerando predições com um menor erro. Poderia à partida ter selecionado variáveis baseado na matriz de correlações acima, mas como não são muitas variáveis, comparadas com o número de observações e como não tenho uma confiança total nas operações transformadas, decidi deixar todo o processo de seleção à regularização.

Para tirar melhor partido do data-set, que é pequeno, mas também para obter repostas menos enviesadas e mais estáveis, vou usar k-fold cross validation sempre que seja necessário avaliar um modelo treinado numa porção de dados não usada para esse treino.

Antes de se poderem tirar conclusões sobre o poder explicativo das variáveis tem que se perceber se as mesmas contêm informação sobre o número de praticantes e como tal se são úteis para o prever. Para tal vou usar um esquema de nested cross validation, em que k-outer_folds são usadas para calcular a performance do modelo e k-inner_folds são usadas em cada fit do outer loop para determinar os hiper-

parâmetros ideais. Este processo foi repetido em 50 trials e os resultados podem ser vistos na **Fig5** em seguida.

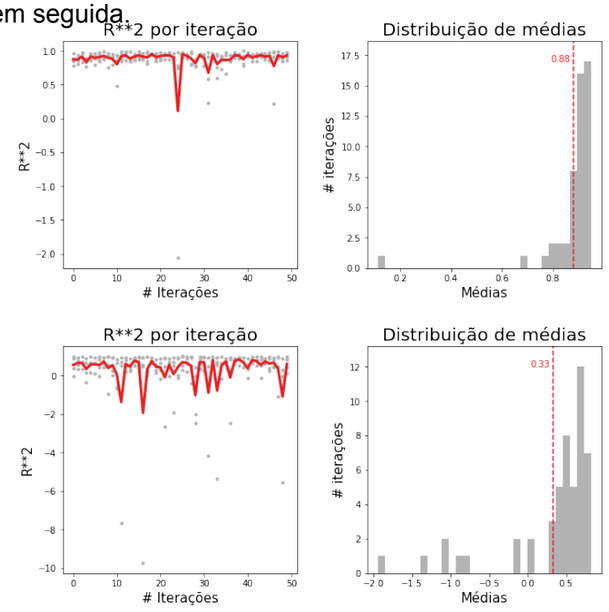

Fig5. Topo: resultados do incluindo a variável *IDH*. Fundo: resultados sem a variável *IDH*. Esquerda: R^2 em cada uma das 4 outer_folds dos 50 trials - bolas cinzentas; R^2 médio em cada trial - linha vermelha. Direita: histograma dos R^2 médios por trial; A linha vermelha representa o valor médio.

Como medida de erro na figura acima, e nas seguintes, usei o coeficiente de determinação (R^2), uma medida que mede quanto melhor um determinado modelo explica a variância de uma variável alvo comparado com a média dessa mesma variável. Assim, um R^2 de 1 significa que as variáveis no nosso modelo explicam na totalidade a variância do alvo e um R^2 de 0 significa que as variáveis do nosso modelo não acrescentam mais informação que a média da variável que estamos a tentar prever. Um R^2 negativo significa que o nosso modelo teve um desempenho pior que a média, o que pode acontecer quando estamos a testar o modelo em dados que não foram utilizados para o treinar.

Na figura 5, acima, podemos constatar que as nossas variáveis explicativas contêm informação que pode ser utilizada para explicar o número de participantes por época. Na maior parte das partições train - test resultantes do processo de cross validation o R^2 foi superior a 0.

As instâncias do modelo que incluem a variável IDH (plots em cima na figura 5) têm uma performance particularmente boa: R^2 médio por trial de 0.88 e mais de 98% dos testes em dados não usados para treinar (círculos cinzentos nos gráficos) com um $R^2 > 0.5$.

Nas instâncias do modelo em que a variável *IDH* foi deixada de fora, por não se tratar de uma variável directamente relacionada com a visibilidade da modalidade, as performances foram inferiores mas ainda assim positivas: R^2 médio por trial de 0.33 e cerca de 65% dos testes em dados não usados para treinar (círculos cinzentos nos gráficos) com um $R^2 > 0.5$.

Agora que tenho confiança que as variáveis explicativas podem ser usadas com sucesso para prever o número de praticantes por época, em dados que não foram usados para determinar os coeficientes atribuídos a cada uma, posso olhar para esses mesmos coeficientes para perceber qual as mais importantes e qual é o seu poder explicativo. Para tal vou treinar um modelo usando os dados todos e cross validation para determinar os melhores hiper-parâmetros: o balanço entre *ridge* e *lasso*, já que estou a usar uma regularização elasticnet; e o parâmetro que determina a força dessa regularização.

Em baixo os coeficientes ordenados, por valor absoluto, associados a cada uma das variáveis incluídas nos modelos:

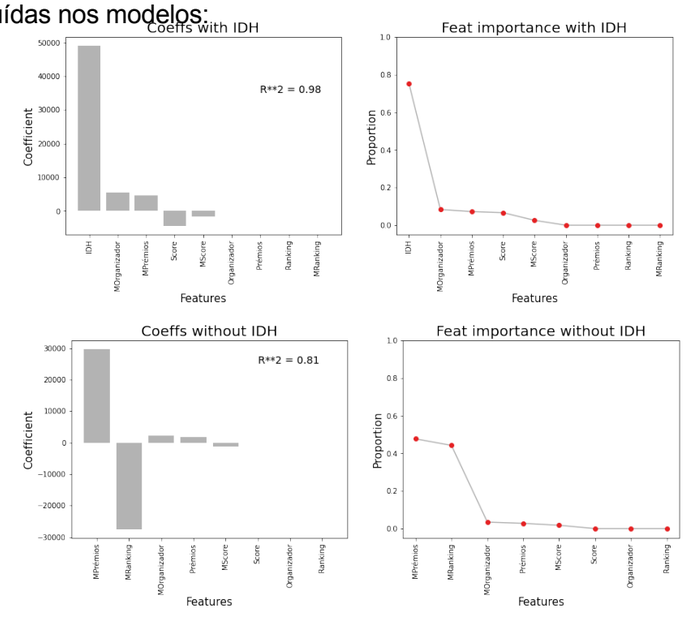

Fig6. Topo: resultados incluindo a variável *IDH*. Fundo: resultados sem a variável *IDH*. Esquerda: Coeficientes calculados pelo modelo para cada variável ordenados por valor absoluto. Direita: importância relativa de cada variável na predição final.

Olhado para o fundo da figura 6, o modelo sem *IDH*, constata-se que um modelo linear construído com as variáveis associadas à visibilidade da modalidade consegue explicar 88% da variância na variável dependente, o número de praticantes por época. É de destacar que não é claro que o limite superior deste valor fosse 100% uma vez que isto assumiria uma relação linear entre as variáveis explicativas e variável alvo, o que não é certa.

As variáveis que dominam em importância a predição do modelo são as *MPrémios*, coef = 29561, e *MRanking*, coef = -27416. Os coeficientes são desta ordem de grandeza porque as variáveis foram escaladas entre 0 - 1 antes de serem passadas à regressão linear. Tal facilita o trabalho do algoritmo em encontrar a solução óptima e, crucialmente, permite que os coeficientes das diversas variáveis possam ser comparados directamente, o que não seria possível se as variáveis mantivesses as suas unidades originais. A interpretação directa destes coeficientes é que a uma deslocação de uma unidade na escala da variável explicativa (neste caso depois de escalada) corresponde uma alteração igual ao coeficiente na variável alvo. Claro que depois da transformação por operada nas variáveis *Ranking* e *Prémios* a noção do que é uma unidade perde clareza. De notar também que *MPrémios* e *MRanking* têm coeficientes com sinais diferentes, devido ao sinal diferente da sua correlação com a variável número de praticantes.

Há outras variáveis com coeficientes diferentes de 0, nomeadamente *MOrganizador*, *Prémios* e *MScore*, mas estes representam apenas 3.4, 2.8 e 1.8% (total aprox 8%), respectivamente, da capacidade do modelo em explicar o número de praticantes. O que contrasta com os cerca de 92% explicados pelas duas variáveis mais importantes.

Chamar apenas a atenção para o facto de o modelo que usa a variável *IDH*, conseguir explicar 98% da variância na variável alvo. Neste caso a variável *IDH* domina completamente a capacidade explicativa do modelo representando 75% desta.

4. Principais conclusões e comentários

A pergunta original que motivou esta investigação foi: *Qual o contributo, no número de praticantes por época, de variáveis relacionadas com a visibilidade da modalidade?* Depois da análise efectuada a resposta que posso dar é que as variáveis associadas à visibilidade que mais contribuem para o número de praticantes por época são variáveis que resultam da transformação das variáveis originais que descrevem os prémios recebidos e o ranking da seleção. Estas duas variáveis juntas explicam mais de 90% da variância que o modelo de regressão linear consegue explicar no número de adeptos por época. Isto está de acordo com os resultados previamente obtidos com a análise de correlações em que estas foram também as variáveis com uma correlação mais forte com a variável alvo.

A noção de variável importante e mesmo o peso atribuído a cada variável são conceitos complexos. Uma variável pode não ser importante por si mas ser importante quando complemento de outros. Por exemplo a variável *MOrganizador* é a segunda com maior peso no modelo que inclui a variável *IDH*, apesar de a sua correlação com o número de praticantes ser a mais baixa. O contraste entre os modelos com e sem *IDH* é também a prova de como a simples introdução de uma variável pode alterar completamente os coeficientes e importância relativa das outras. Os resultados aqui apresentados são assim sólidos no contexto das variáveis utilizadas, mas podiam facilmente mudar com a introdução de novas variáveis ou até de novas observações com informação diferente. O dados disponíveis eram relativamente poucos.

Acho que teria sido interessante juntar a estas variáveis que dissessem directamente respeito à visibilidade da modalidade: tempo de antena, campanhas publicitárias, campanhas de angariação, investimento em publicidade, presença nas redes sociais, etc. Afinal são estas as intermediárias entre o desempenho desportivo e a visibilidade em si.

As alterações às variáveis foram feitas baseadas em heurísticas simples e não verificadas. Com mais tempo podiam ter-se feito de uma maneira mais orientada efectuando-se, por exemplo, grelhas de busca para perceber qual a deslocação temporal ideal entre determinadas variáveis e a variação no número de adeptos, ou qual a janela ideal para as médias rolantes. Mais do que truques para melhorar a performance dos modelos estes parâmetros podem fornecer informações importantes sobre a maneira com as variáveis explicativas se relacionam com a variável alvo.