Investingando a exclusão de Pontos Influentes

Alisson Rosa e Vítor Pereira

Sumário

1 Análise de Influência **1**
1.1 Exclusão de poucas observações 1
1.2 Exclusão de muitas observações 4
1.3 Testes 7
1.4 Critério de seleção de modelos 8

Referências **8**

1 Análise de Influência

Nesse anexo verificaremos o que aconteceria se tomassemos a decisão de remover os pontos aberrantes, assim iremos desenvolver duas análises sobre a existência de observações atípicas, isto é, vamos considerar dois eventos da exclusão de pontos que exercem peso desproporcional no modelo de Regressão Logística, assim sucedendo com análise de pontos de avalanca, distância de cook, dffits e envelope simulado.

1.1 Exclusão de poucas observações

Nessa seção realizaremos a análise de influência para a remoção de poucas observações, que podem ser pontos influentes, apenas três as observações 262,430 e 415 (sendo os Pokémons conhecidos como: Blissey, Deoxys Attack, Regirock).

```
## [05:22:52] WARNING: amalgamation/../src/learner.cc:1115: Starting in XGBoost 1.3.0, the default evaluatio
```

1.1.1 Matrizes de confusões para os modelos propostos

Não devemos somente avaliar os gráficos de influência, mas também sua previsão, então tem-se que uma medida interessante é a matriz de confusão que pode ser vista como uma tabela que possui os valores reais cruzados com os valores preditos, vejamos para os 3 modelos ajustados como a matriz de confusão fica para os dados de teste:

Tabela 1: Matriz de confusão para os dados de teste no modelo de Regressão logística

	Predição		
Legendary	Não	Sim	Total
Não	186	1	187
Sim	6	7	13
Total	192	8	200

Tabela 2: Matriz de confusão para os dados de teste no modelo de Random Forest

	Predição		
Legendary	Não	Sim	Total
Não	187	0	187
Sim	9	4	13
Total	196	4	200

Tabela 3: Matriz de confusão para os dados de teste no modelo de XgBoost

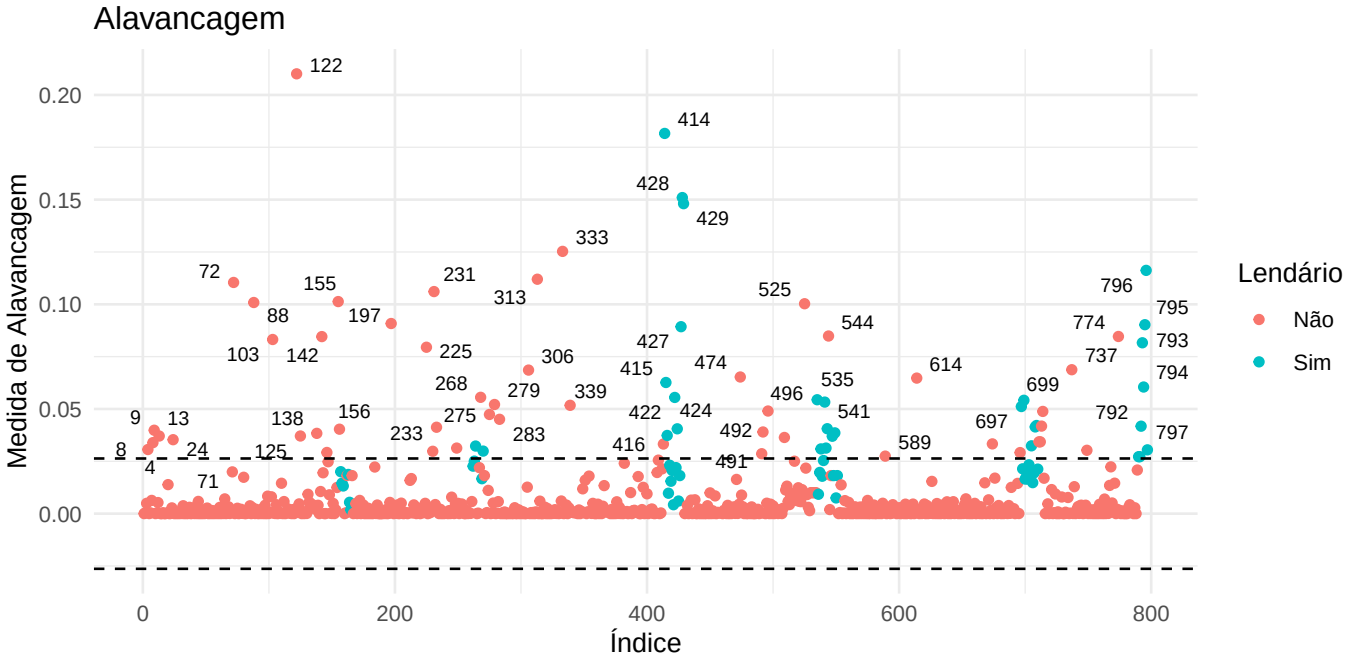
	Predição		
Legendary	Não	Sim	Total
Não	186	1	187
Sim	6	7	13
Total	192	8	200

Não é perceptível uma diferença relevante das predições dos modelos com exclusão dos pontos aberrantes para a predição dos modelos ajustados no trabalho principal, pois vemos uma piora na Matriz de confusão da Random Forest e uma melhora no modelo do XGBoost.

1.1.2 Análise de Dignóstico

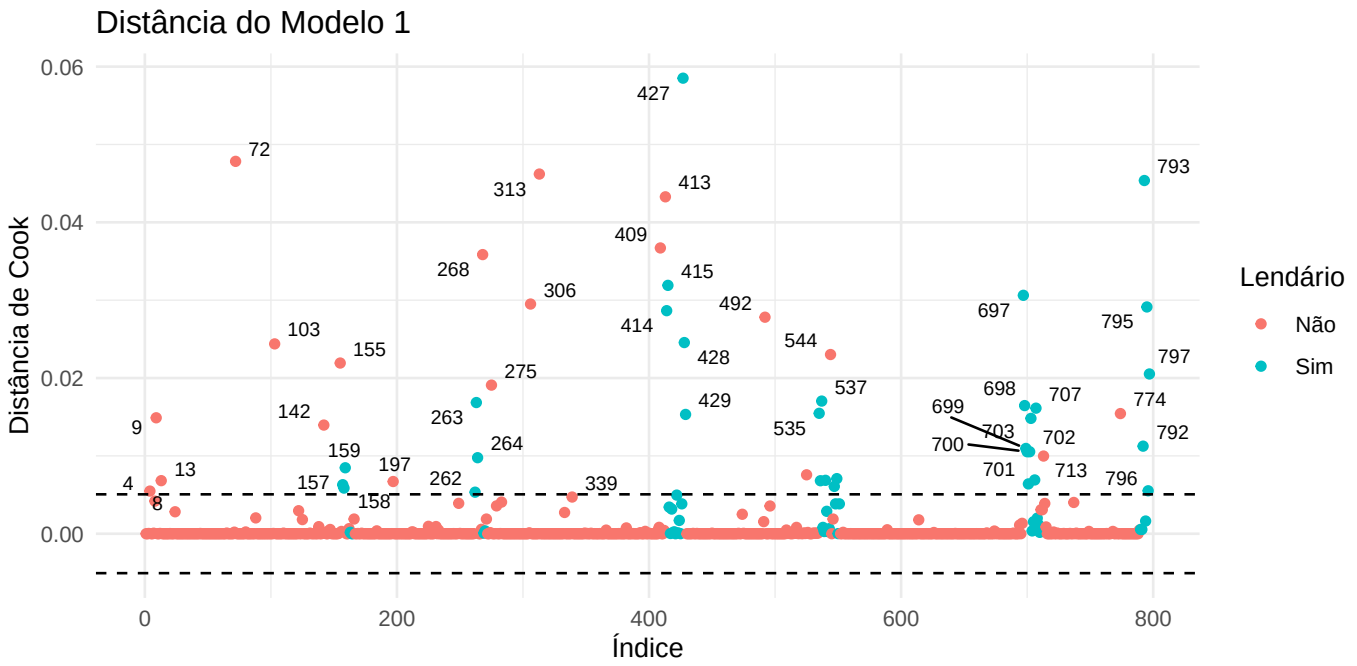
Vamos nessa seção avaliar a existência de pontos influentes com a exclusão dos pontos influentes anteriores no modelo de Regressão logística.

1.1.2.1 Alavancagem Nessa seção veremos as medidas de alavancagem, que informam se uma observação é discrepante em termos de covariável, ou seja, utilizando os resíduos busca medir a discrepância entre o valor observado e o valor ajustado.



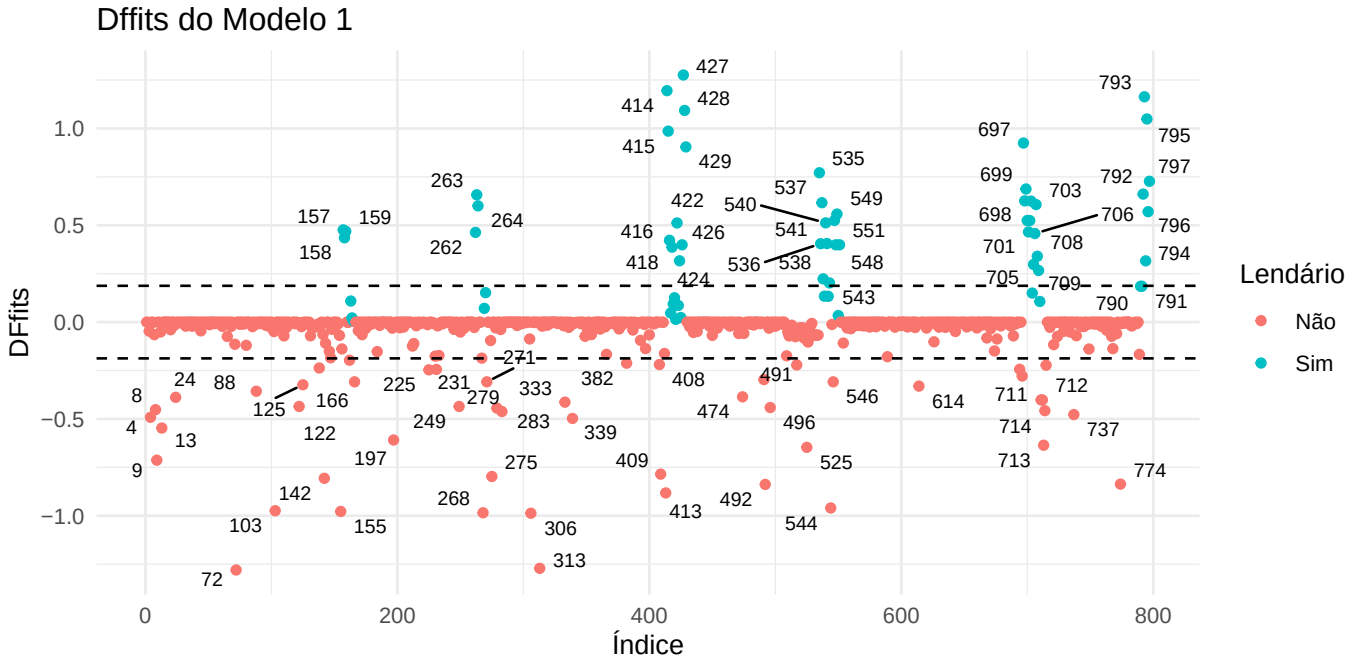
Podemos perceber que continuam existindo pontos fora do intervalo traçado para alavancagem.

1.1.2.2 Distância de cook Tem-se também a distância de cook, que fornece a influência da observação i sobre todos os n valores ajustados.



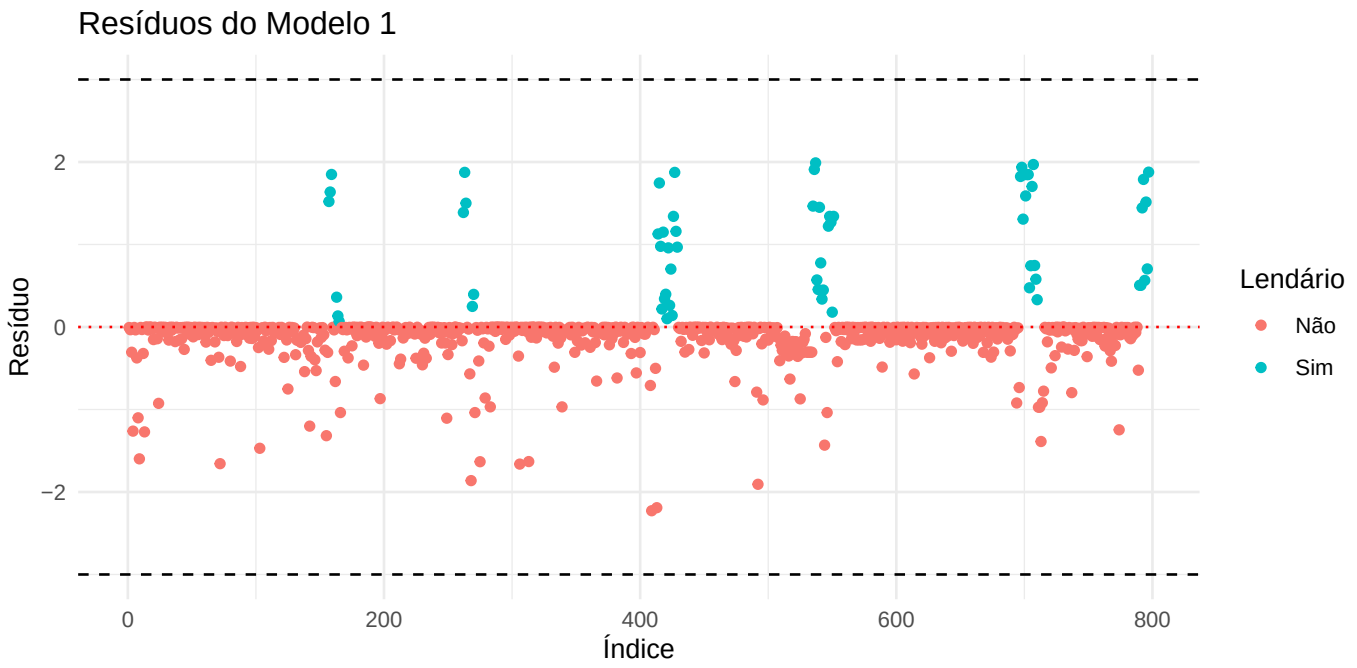
A existência de observações possivelmente discrepantes, ainda não foi reduzida visualmente.

1.1.2.3 Dffits No diagnóstico dffits, que informam o grau de influência que a observação i tem sobre o valor seu próprio valor ajustado $\hat{y}_i$, percebe-se que:



Não observa-se uma mudança perceptível quanto ao achatamento do gráfico.

1.1.2.4 Resíduos O gráfico de resíduos também é importante para verificarmos visualmente a média dos resíduos e se existe algum valor fora do limite de 3 desvios padrões, pois esses possui baixíssima probabilidade de serem observados, no gráfico abaixo verificamos que todos os estados estão dentro dos limites:



Para os resíduos, os pontos continuam dentro dos limites desenvolvidos.

1.1.2.5 Envelope Simulado E por último o envelope simulado, que fornece um vislumbre se a distribuição é adequada para o ajuste

É visível a melhora comparado ao envelope simulado do modelo principal, mas ainda temos observações fora das bandas simuladas.

1.2 Exclusão de muitas observações

Nessa seção consideraremos o segundo evento, assim realizaremos a análise de influência para a remoção de muitas observações, que podem ser pontos influentes, totalizando um total de 24 pokémons removidos, desses 10 são lendários, ou seja, 15.385% de todos os lendários, reduzindo ainda mais a condição rara.

```
## [05:23:14] WARNING: amalgamation/../src/learner.cc:1115: Starting in XGBoost 1.3.0, the default evaluatio
```

1.2.1 Matrizes de confusões para os modelos propostos

Não devemos somente avaliar os gráficos de influência, mas também sua previsão, então tem-se que uma medida interessante é a matriz de confusão que pode ser vista como uma tabela que possui os valores reais cruzados com os valores preditos, vejamos para os 3 modelos ajustados como a matriz de confusão fica para os dados de teste:

Tabela 4: Matriz de confusão para os dados de teste no modelo de Regressão logística

	Predição		
Legendary	Não	Sim	Total
Não	180	1	181
Sim	2	11	13
Total	182	12	194

Tabela 5: Matriz de confusão para os dados de teste no modelo de Random Forest

	Predição		
Legendary	Não	Sim	Total
Não	181	0	181
Sim	6	7	13
Total	187	7	194

Tabela 6: Matriz de confusão para os dados de teste no modelo de XgBoost

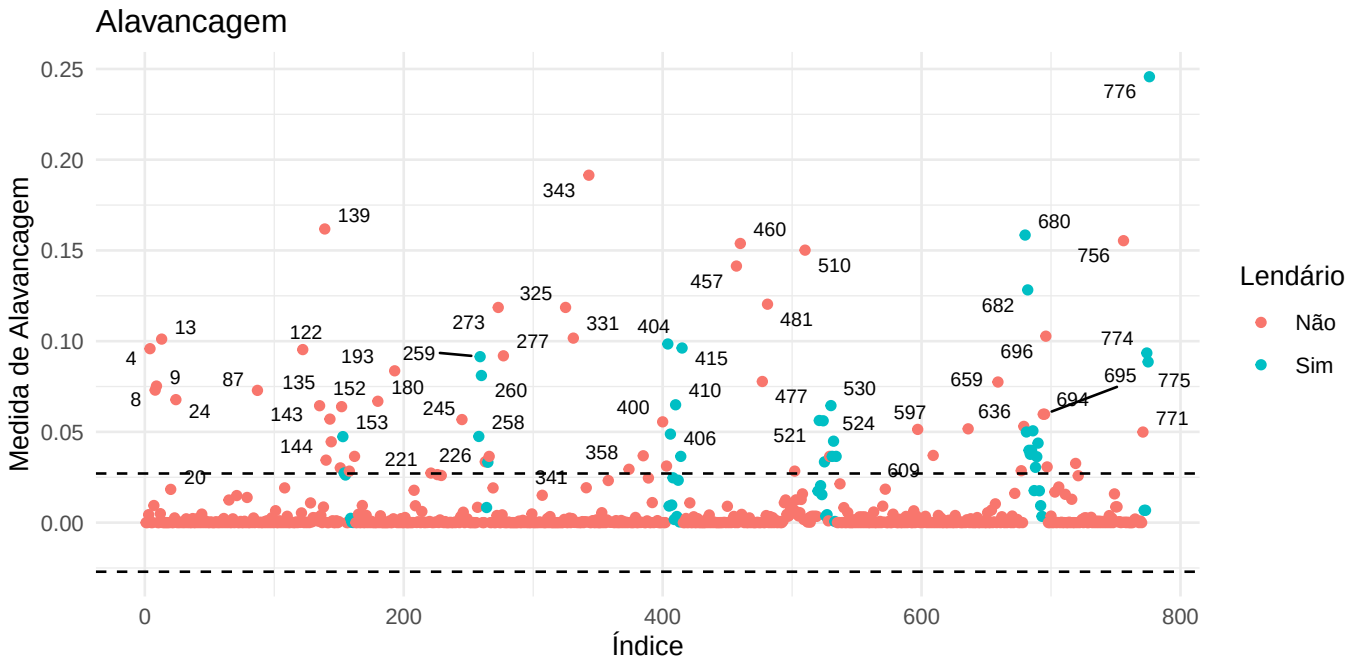
	Predição		
Legendary	Não	Sim	Total
Não	181	0	181
Sim	1	12	13
Total	182	12	194

É perceptível uma diferença relevante das predições dos modelos com exclusão de muitos pontos aberrantes para a predição dos modelos ajustados no trabalho principal, é notada uma melhora da sensibilidade em todos os modelos.

1.2.2 Análise de Dignóstico

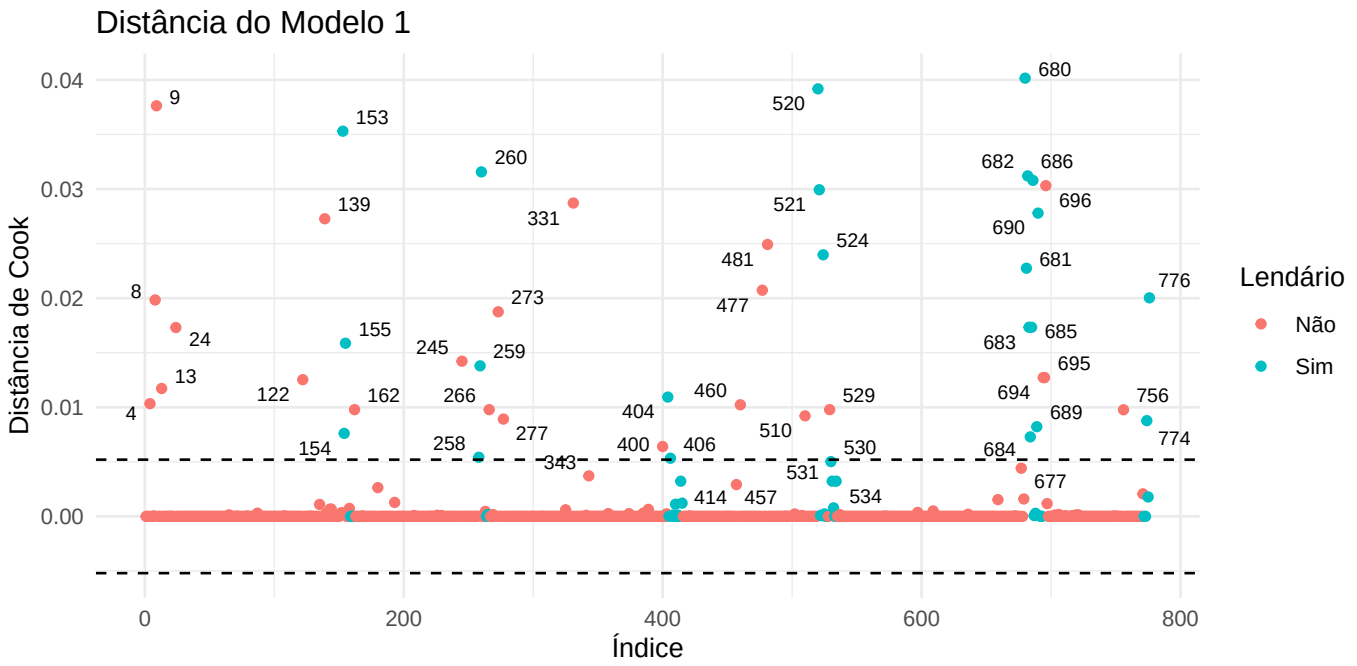
Vamos nessa seção avaliar a existência de pontos influentes com a exclusão dos pontos influentes anteriores no modelo de Regressão logística.

1.2.2.1 Alavancagem Nessa seção veremos as medidas de alavancagem, que informam se uma observação é discrepante em termos de covariável, ou seja, utilizando os resíduos busca medir a discrepância entre o valor observado e o valor ajustado.



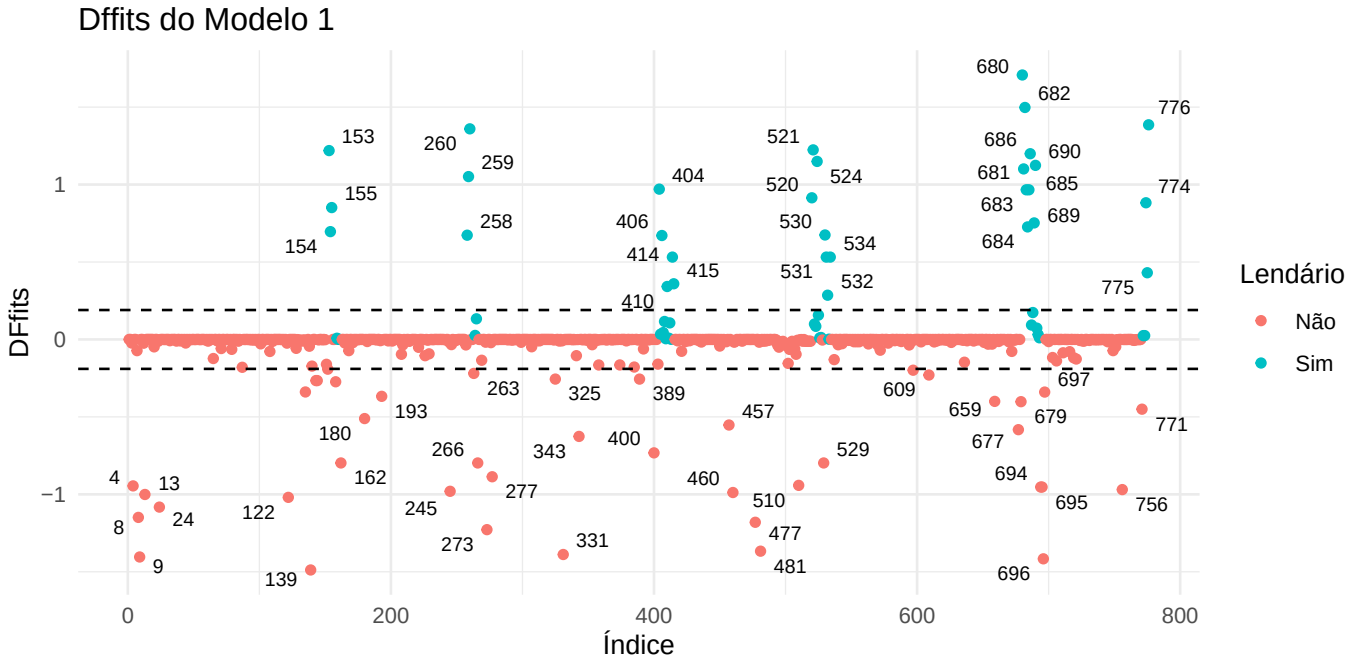
Ainda continuamos a perceber pontos fora do intervalo traçado para alavancagem.

1.2.2.2 Distância de cook Tem-se também a distância de cook, que fornece a influência da observação i sobre todos os n valores ajustados.



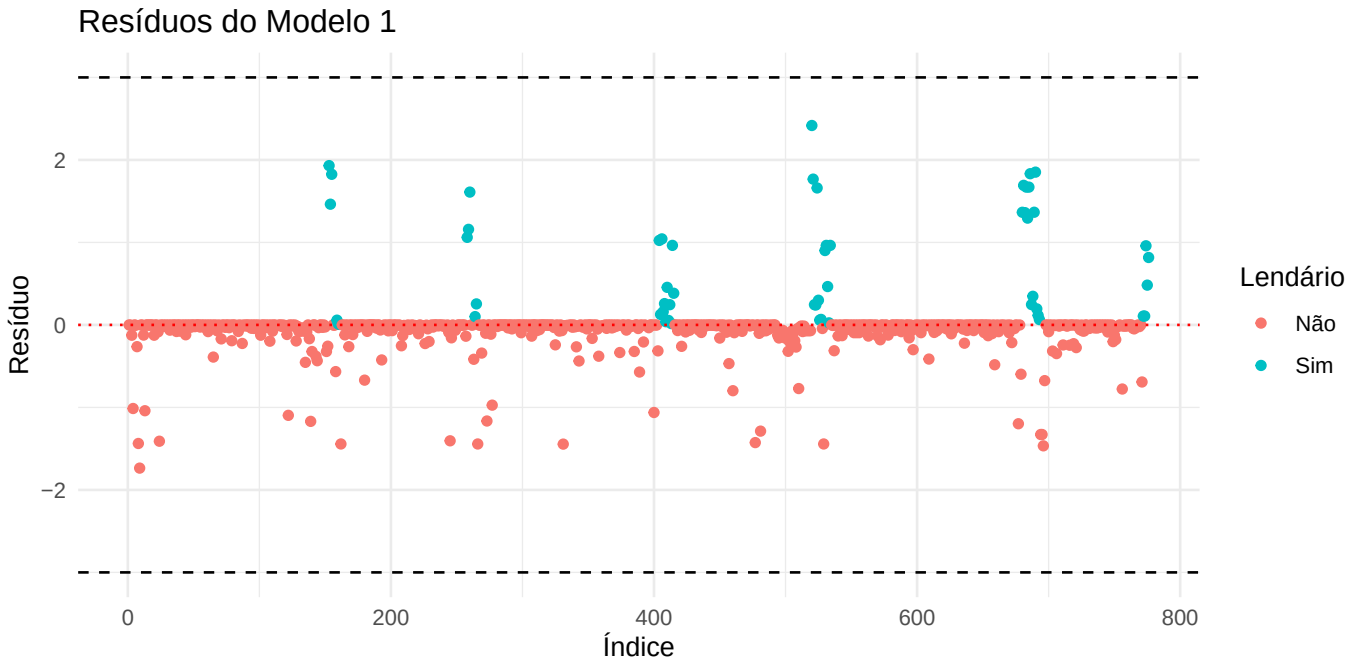
A existência de observações possivelmente discrepantes, ainda não foi reduzida visualmente.

1.2.2.3 Dffits No diagnóstico dffits, que informam o grau de influência que a observação i tem sobre o valor seu próprio valor ajustado $\hat{y}_i$, percebe-se que:



Observa-se que o achatamento do gráfico ainda continua nas mesmas proporções.

1.2.2.4 Resíduos O gráfico de resíduos também é importante para verificarmos visualmente a média dos resíduos e se existe algum valor fora do limite de 3 desvios padrões, pois esses possui baixíssima probabilidade de serem observados, no gráfico abaixo verificamos que todos os estados estão dentro dos limites:



Para os resíduos, os pontos continuam dentro dos limites desenvolvidos.

1.2.2.5 Envelope Simulado E por último o envelope simulado, que fornece um vislumbre se a distribuição é adequada para o ajuste

A melhora comparado ao envelope simulado do modelo principal e ao envelope simulado anterior é notável, mas com todas as remoções ainda não tem-se observações fora das bandas simuladas.

1.3 Testes

Verificando a significância das covariáveis para os modelos com remoções de pontos influentes

Tabela 7: Estatísticas do Modelo com poucas remoções

term	estimate	std.error	statistic	p.value
(Intercept)	-22.740	2.589	-8.78	0.000
HP	0.043	0.011	4.09	0.000
Attack	0.015	0.007	2.18	0.030
Defense	0.030	0.009	3.27	0.001
‘Sp. Atk‘	0.035	0.007	4.73	0.000
‘Sp. Def‘	0.049	0.010	5.07	0.000
Speed	0.052	0.010	5.28	0.000

Tabela 8: Estatísticas do Modelo com muitas remoções

term	estimate	std.error	statistic	p.value
(Intercept)	-37.035	5.673	-6.53	0.000
HP	0.101	0.022	4.61	0.000
Attack	0.027	0.010	2.67	0.008
Defense	0.046	0.015	3.07	0.002
‘Sp. Atk‘	0.051	0.011	4.40	0.000
‘Sp. Def‘	0.046	0.015	3.08	0.002
Speed	0.105	0.020	5.35	0.000

Então, continua-se com todas as covariáveis significativas para o modelo, as principais diferenças entre os três modelos estão associadas ao Intercepto e a covariável Speed.

1.4 Critério de seleção de modelos

Avaliando quanto aos critérios de seleção, temos:

Tabela 9: Critérios de Seleção do Modelo para a Regressão Logistica

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
441	796	-83.6	181	214	167.2	790	797
397	775	-48.9	112	144	97.9	769	776

Assim o modelo com mais remoções tem os critérios de AIC e BIC melhores, logo podemos reafirmar a decisão tomada anteriormente, pois estaríamos sacrificando uma grande parte da classificação objetivo para buscar o melhor ajuste possível para a análise de influência, mas ainda sem conseguir efetivamente, pois muitas observações estão fora dos limites e principalmente ainda contêm pontos fora das bandas simuladas do envelope, então mesmo com renúncia de uma maior quantidade para a classificação rara ainda não conseguimos atingir o objetivo da análise de influência.

Referências

[1] Casella G, Berger RL. Inferência estatística. Cengage Learning; 2021.

[2] R Core Team. R: A language and environment for statistical computing [Internet]. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available from: https://www.R-project.org/.

[3] Pokémon wiki [Internet]. Fandom. Available from: https://pokemon.fandom.com/wiki/Pok%C3%A9mon_Wiki.

[4] Pokémon [Internet]. Wikipedia. Wikimedia Foundation; 2022. Available from: https://en.wikipedia.org/wiki/Pok%C3%A9mon.