Explorando a mística de Pokémon: Caracterização de Pokémons Lendários

Alisson Rosa e Vítor Pereira

Resumo

Pokémon são criaturas que vivem em todos os lugares, livres na natureza ou com os humanos, cada Pokémom tem seu tipo, pontos fortes e fracos. Com isso o objetivo desse trabalho é analisar suas estatísticas, desenvolvendo gráficos e tabelas e também construindo um modelo que dados as características do Pokémon ele irá nos fornecer uma predição se o Pokémon é lendário ou não.

Sumário

Tabela 1: Média dos atributos entre as classificações

Legendary	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	n
Não	67.2	75.7	71.6	68.5	68.9	65.5	735
Sim	92.7	116.7	99.7	122.2	105.9	100.2	65

1 Análise Preditiva

Tabela 2: Matriz de confusão para os dados de teste no modelo de Regressão logística

	Predição		
Legendary	1	0	Total
1	9	9	18
0	1	181	182
Total	10	190	200

comentários

Da matriz de confusão podemos derivar as seguintes métricas:

- Valor predito positivo (`ppv`): Que é definido como sendo a proporção de predições positivas que foram corretamentes previstas
- Valor predito negativo (`npv`): Por definição é a proporção de predições negativas que foram corretamentes previstas
- Sensibilidade (`sens`): É a proporção de previsões corretas dos casos positivos

- Especificidade (`spec`): É a proporção de previsões corretas dos casos negativos

E para os dados de teste temos:

Tabela 3: Métricas nos dados de teste

Modelo	sens	spec	ppv	npv	roc_auc
Regressão Logistica	0.5	0.995	0.9	0.953	0.984

Assim vemos pela tabela 7 que o modelo que maximizou a sensibilidade é Regressão Logistica

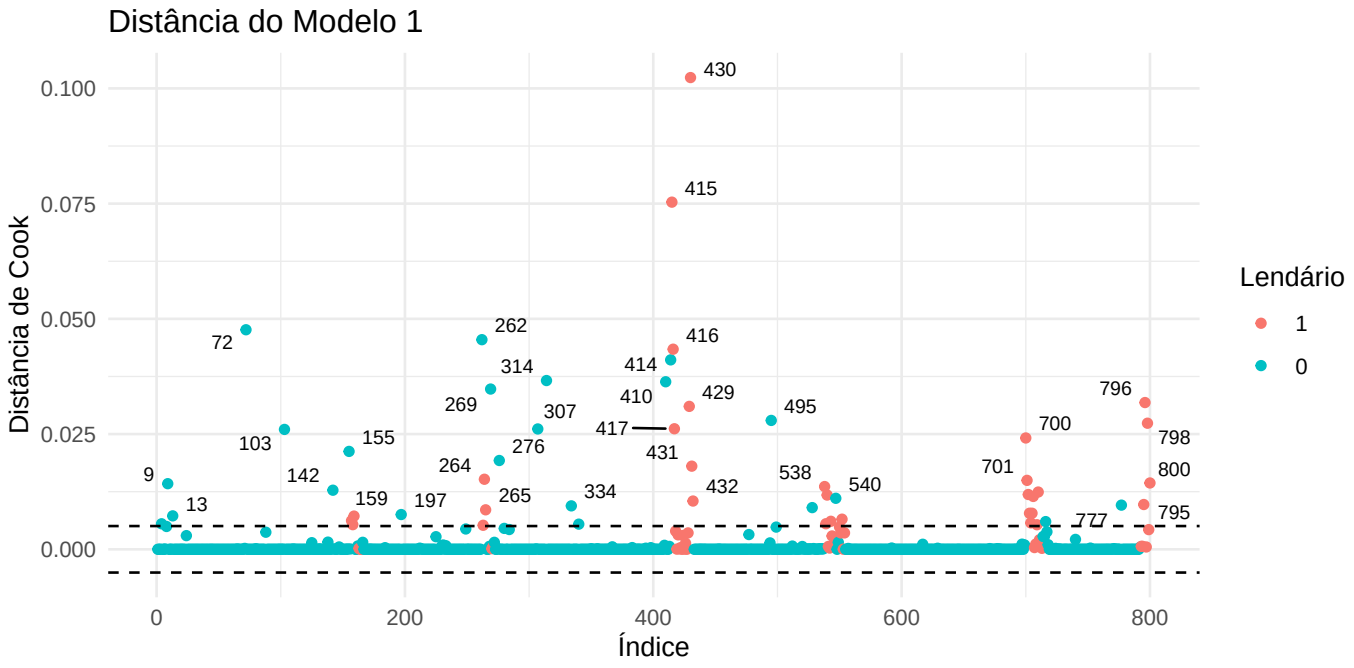
2 Análise Inferencial

2.1 Análise de Dignóstico

Vamos nessa seção avaliar a existência de pontos influentes no modelo de Regressão logística

2.1.1 Distância de cook

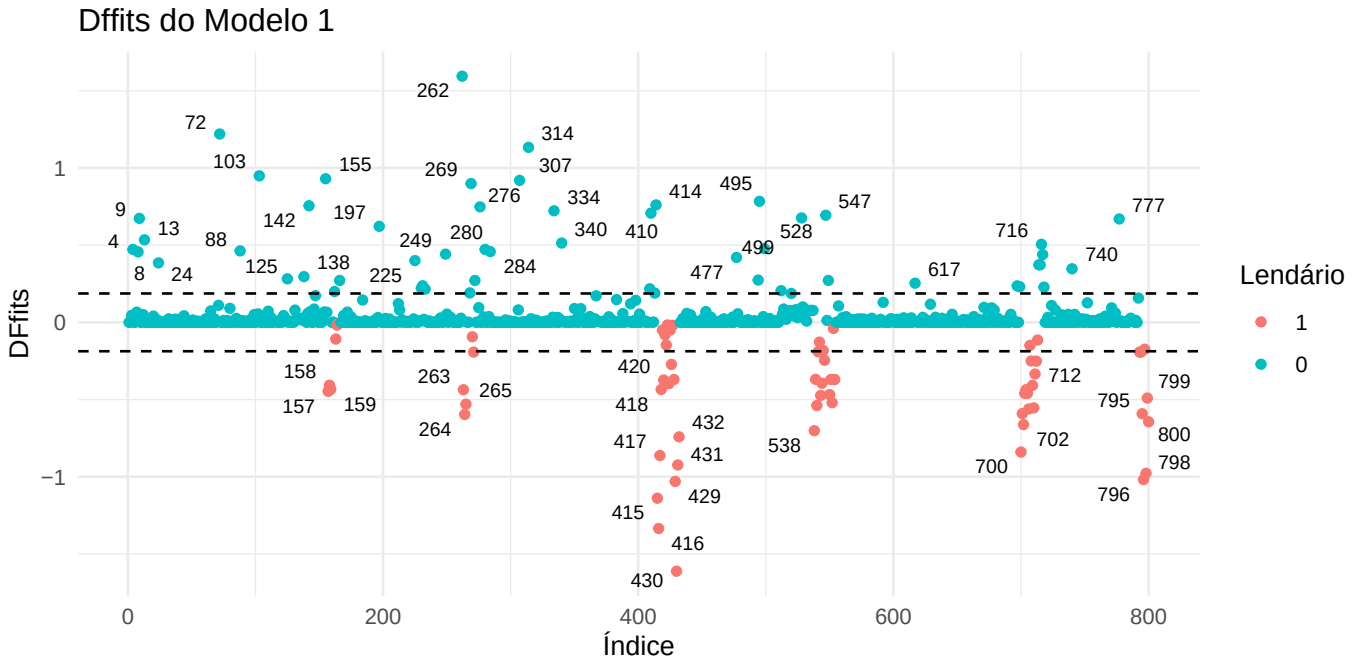
Tem-se também a distância de cook, que fornece a influência da observação i sobre todos os n valores ajustados,



Pontos ppossivelmente influentes => 430 e 415

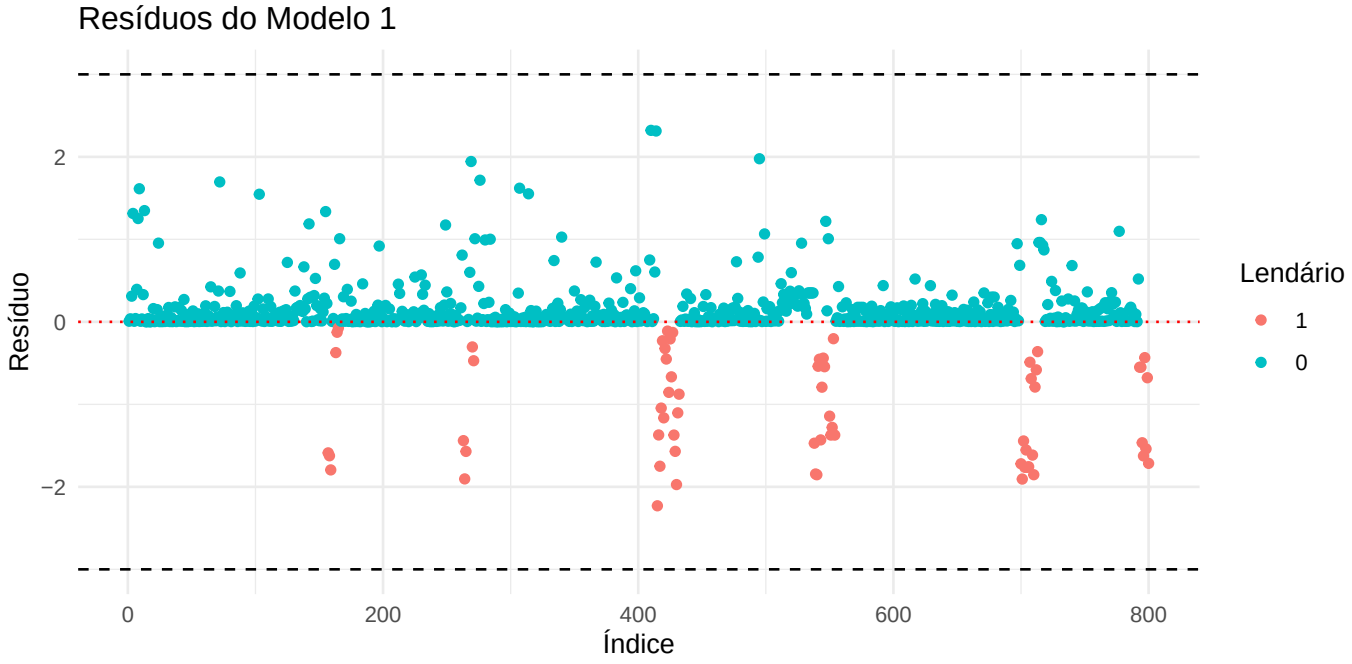
2.1.2 Dffits

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:



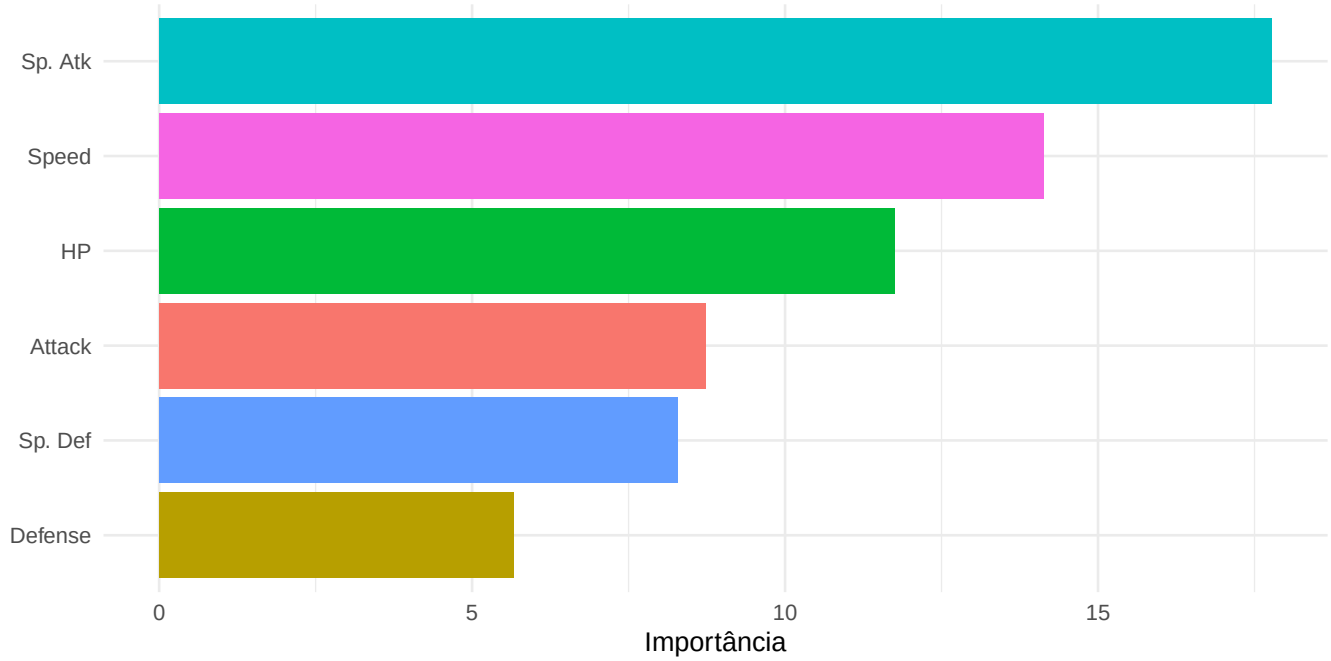
Pontos ppossivelmente influentes => 262, 72, 314, 430, 416 e 415

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:



E por último o envelope simulado, que fornce um vislumbre se a distribuição é adequada para o ajuste

3 Deletando 415 e 430




```
## [07:33:12] WARNING: amalgamation/../src/learner.cc:1115: Starting in XGBoost 1.3.0, the default evaluatio
```

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	182	1	183
Sim	6	9	15
Total	188	10	198

comentários

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

	Predição		
Legendary	Não	Sim	Total
Não	181	2	183
Sim	5	10	15
Total	186	12	198

comentários

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

	Predição		
Legendary	Não	Sim	Total
Não	181	2	183
Sim	5	10	15
Total	186	12	198

comentários

O gráfico ref fornece o vislumbre de como as métricas se comportam para os 3 modelos ajustados na validação cruzada

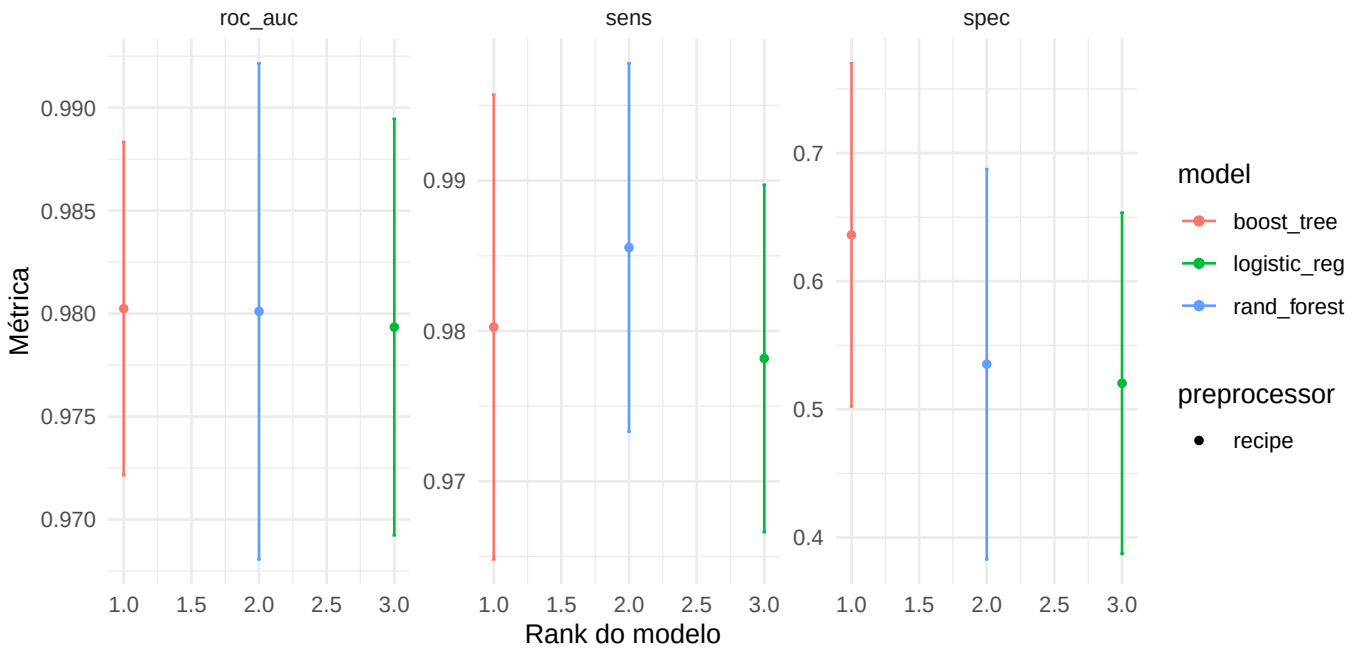


Figura 1: Métricas dos modelos na validação cruzada

E para os dados de teste temos:

Tabela 7: Métricas nos dados de teste

Modelo	sens	spec	ppv	npv	roc_auc
Regressão Logistica	0.995	0.600	0.968	0.900	0.985
Random Forest	0.989	0.667	0.973	0.833	0.984
xgboost	0.989	0.667	0.973	0.833	0.954

Assim vemos pela tabela 7 que o modelo que maximizou a sensibilidade é Regressão Logistica

4 Análise Inferencial

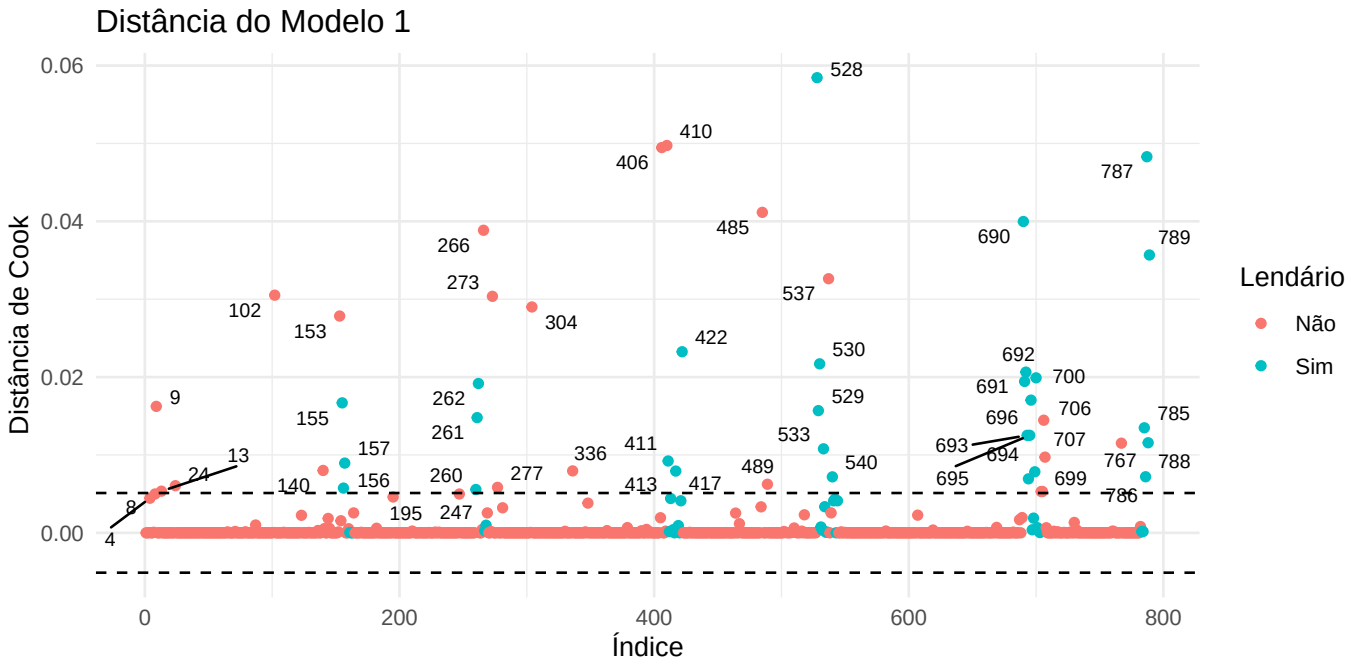
4.1 Análise de Dignóstico

Vamos nessa seção avaliar a existência de pontos influentes no modelo de Regressão logística

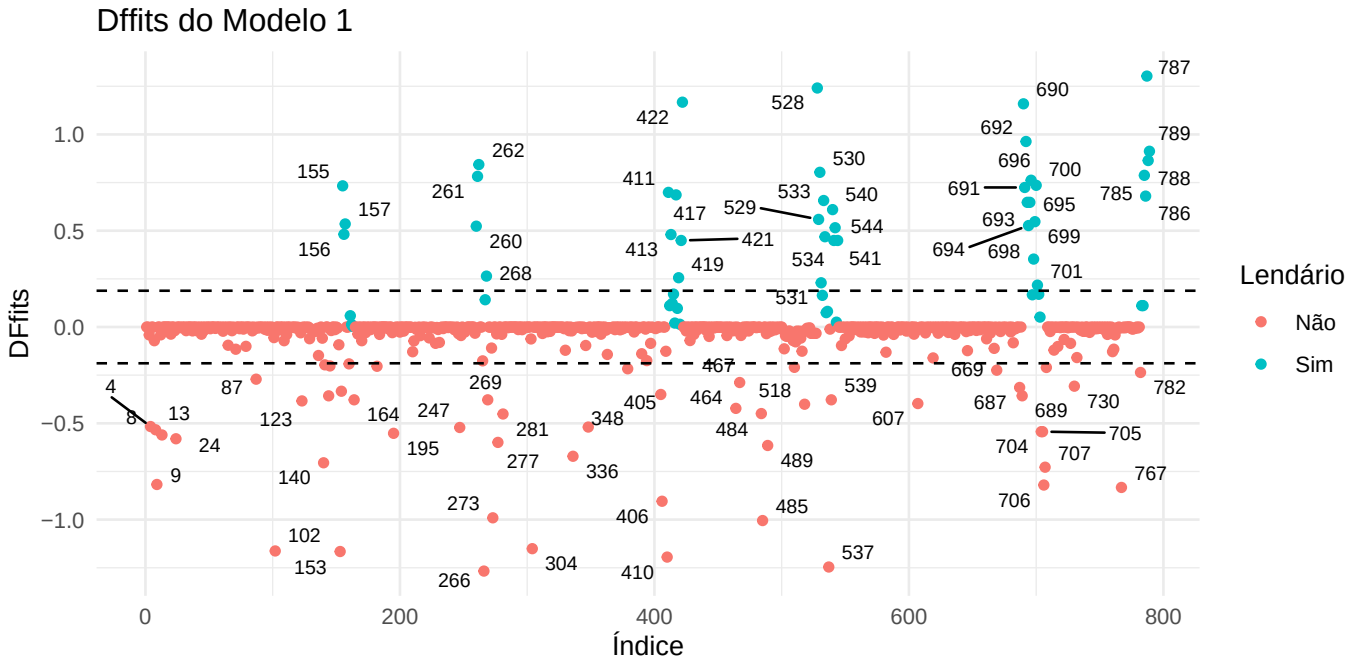
‘

4.1.1 Distância de cook

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

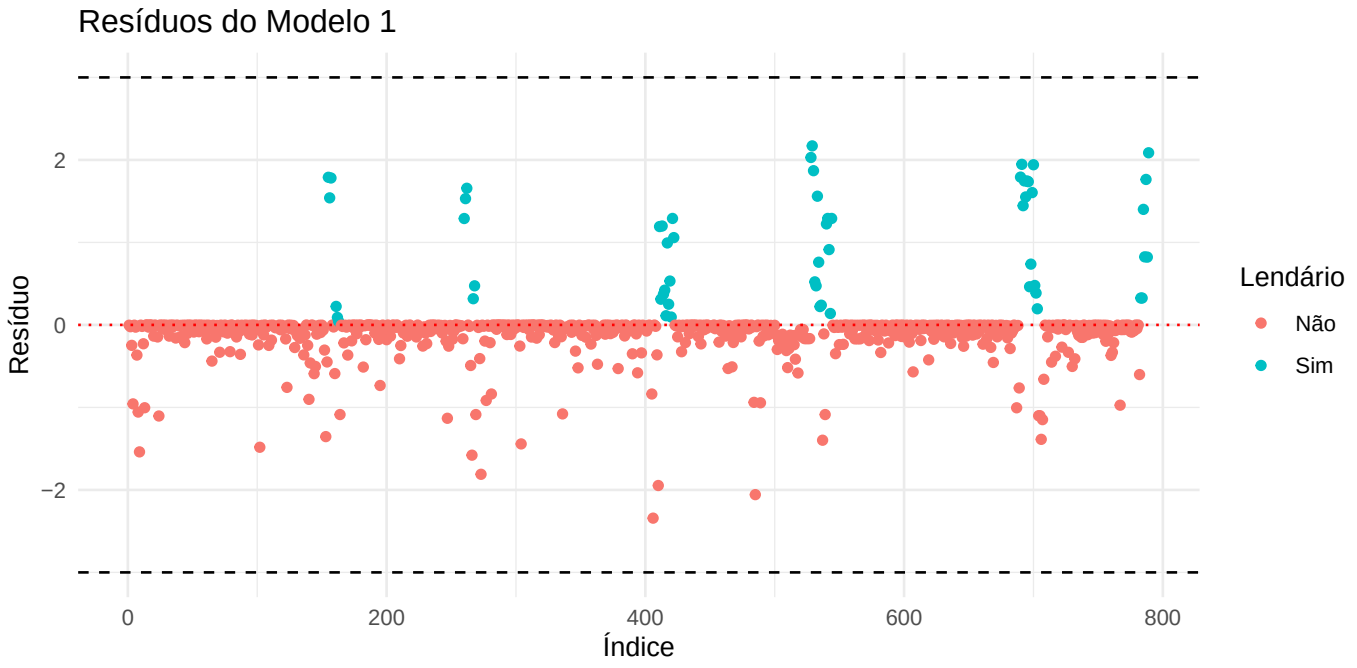


Pontos ppossivelmente influentes => 528,410,406, 787 depois 266, 485, 690 ### Dffits No diagnóstico dffits, que informam o grau de influência que a observação i tem sobre o valor seu próprio valor ajustado $\hat{y}_i$, percebe-se que:



Pontos ppossivelmente influentes => 422, 528, 690, 692, 787, 102, 153, 366,410,537

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:



E por último o envelope simulado, que fornce um vislumbre se a distribuição é adequada para o ajuste

[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.