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

1 Introdução

Há mais de 20 anos, crianças do mundo inteiro vêm descobrindo o mundo encantado de Pokémon e muitas delas se tornam fãs para a vida toda. Hoje, a família de produtos Pokémon inclui videogames, o jogo de cartas Pokémon Estampas Ilustradas, uma série de animação, filmes, brinquedos, livros e muito mais, mas afinal que são Pokémons? Pokémons são criaturas fictícias que pertencem ao universo da série de mesmo nome - Pokémon, são semelhantes a animais do mundo real, podendo viver em bandos ou individualmente, mas também podem ser inspirados em objetos inanimados como velas, sorvetes, chaveiro e outros instrumentos. Originalmente, a série foi criada como um jogo de videogame e, com a sua popularização, se espalhou para diversos outros formatos, como séries de TV, filmes e livros.

A palavra Pokémon é a contração de duas palavras em inglês: pocket, que significa bolso; e monster, que significa monstro. Assim, um Pokémon é um "monstro de bolso", na tradução literal, além de ser uma contração esse seria o nome original da série, devido ao lugar onde os Pokémons são armazenados: as pokébolas, uma espécie de bola pequena para pode-los transportar com mais facilidade, sendo essas basicamente suas casas. Assim as criaturas poderiam descansar após suas batalhas, sendo essa sua principal função explorada no universo Pokémon, em que os monstrinhos lutam de acordo com habilidades da sua tipagem (Fogo, Água, Planta, Pedra, Elétrico, Voador, Lutador, Psíquico, Fantasma, entre outros.).

2 Análise Descritiva

Cada Pokémon tem seus próprios atributos, como HP (Vida), Attack (Ataque), Defense (Defesa), Speed (Velocidade) e outros mais especificos como:

- **Generation** (geração): Uma Geração em Pokémon é um grupo de jogos separados de acordo com os Pokémon que estão incluídos nela. Cada geração possui novos Pokémon, ataques e habilidades que não existem nas gerações anteriores. Aqui portanto cada Pokémon tera sua respectiva geração, sendo tratata como uma variável de fator.
- **Type** (Tipo): São classificações a que estão submetidos todos os Pokémon e técnicas (movimentos). A partir dos tipos, além de ser possível conhecer um pouco mais a natureza de cada Pokémon, é possível também elaborar estratégias de batalha. Isso porque cada tipo tem vantagens e desvantagens sobre outros tipos. Cada Pokémon pode pertencer a até dois tipos, sendo o primeiro deles o primário (Type 1) e o outro, o secundário (Type 2). Por outro lado, cada movimento tem só um tipo. Um Pokémon pode ter até quatro movimentos, mas elas não precisam ser do mesmo tipo que a criatura.
- **Special Attacks** (Sp. $< >$) : Ataques Especiais são movimentos que dão mais dano do que os anteriores, porém possuem um limitador de uso em forma de barra que deve ser carregada. Assim essa variável é dividida em **Sp. Atk** que é a força do ataque especial e **Sp. Def** que é a defesa do ataque especial.
- **Legendary** (Lendário): Pokémon Lendário (Inglês: Legendary Pokémon) é a denominação dada a uma espécie de Pokémon altamente poderosa, raríssima ou, em alguns casos, até mesmo de um único indivíduo, da qual muito se fala em lendas e mitos no mundo Pokémon, e cuja aparição é extremamente rara. Na seção de modelagem utilizaremos como variável a ser predita o Pokémon ser lendário ou não.

2.0.1 Constraste de Atributos

Nessa subseção vamos vislumbrar os atributos dos Pokémon constrastando entre os lendários e não lendários. Primeiro vejamos a média dos atributos dentre as classificações

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

Assim, como esperado os pokémom lendários possuem atributos superiores (na média) do que os não lendários, note que a força do ataque especial dos Pokémon lendários é 1.78 vezes maior que os não lendários. Para conjecturar a cerca dos atributos dos Pokémon, podemos averiguar os valores pormenorizados na Tabela 2.

Tabela 2: Medidas para resumir as variáveis númericas

variable	média	mediana	desvio padrão	min	max
HP	69.3	65	25.5	1	255
Attack	79.0	75	32.5	5	190
Defense	73.8	70	31.2	5	230
Sp. Atk	72.8	65	32.7	10	194
Sp. Def	71.9	70	27.8	20	230
Speed	68.3	65	29.1	5	180

Na Tabela 2 tem-se os valores da média, mediana, desvio padrão, mínimo e máximo, as medidas para média e mediana são consideralvemente próximas, em torno de 70, com apenas a variável Attack com média 79.001. Conjecturando sobre o desvio padrão é plausível afirmar que que as médias para o Attack e Sp. Atk são as menos informativas, dado que, as amplitudes são menores que HP, Defense e Sp. Def, são os atributos que possuem maior desvio padrão. No atributo HP tem-se uma conjuntura peculiar, o fato de seu mínimo ser 1, o que dá-se pelo Pokémon Shedinja, quem tem um conjunto de habilidades especiais, que podem ser lidas aqui.

2.0.2 Tipos e classificação

Vamos aqui estudar a quantidade de tipos por classificação dos Pokémons. A Figura 1 fornece um vislumbre

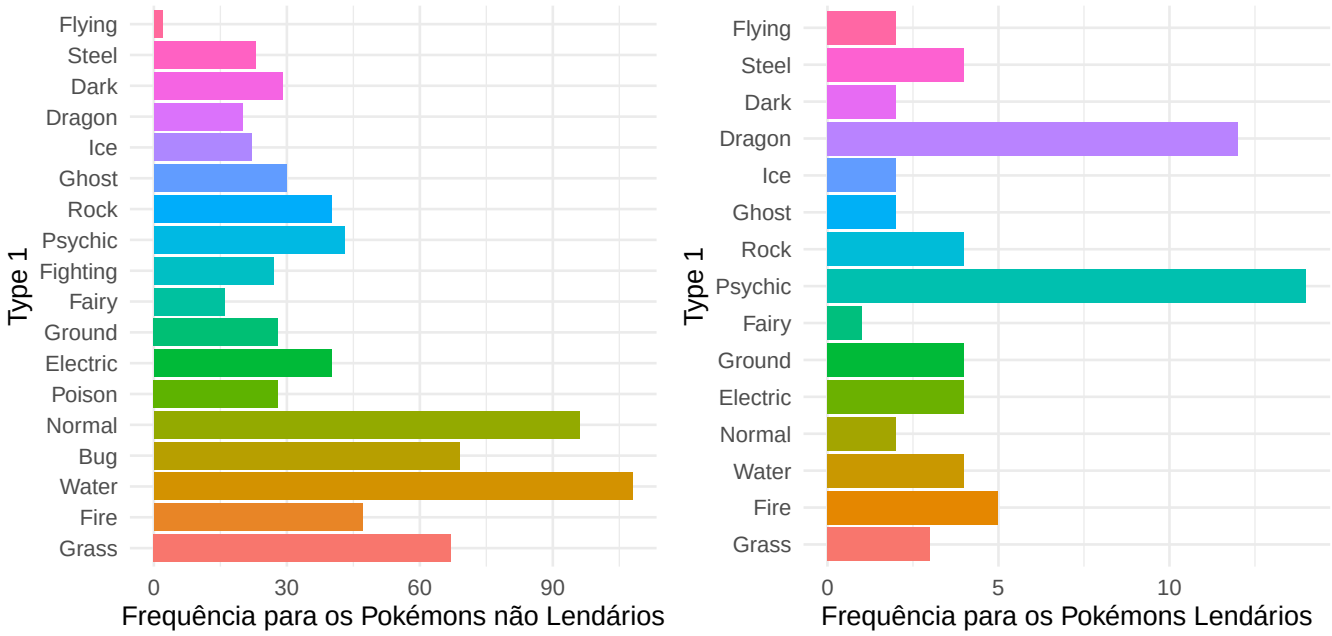

Figura 1: Frequência de Pokémons Lendários e Não Lendários pelo Tipo

Com a Figura 1 nota-se nos Pokémons não lendários que os tipos que se destacam numericamente são Water (água) e Normal, enquanto para os Pokémons lendários se destacam Psychic (psíquico) e Dragon (dragão). Com a Figura 2 podemos ver a quantidade dos tipos secundários dos Pokémons.

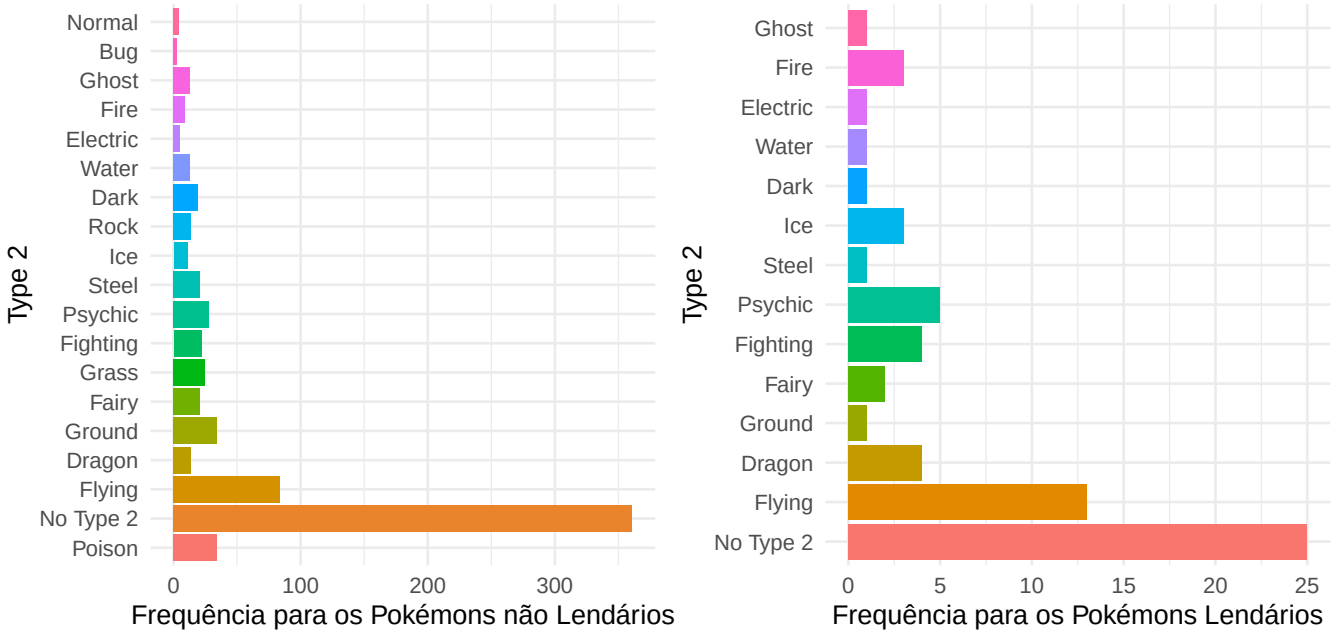

Figura 2: Frequência de Pokémons Lendários e Não Lendários pelo Tipo secundário

Identifica-se pela Figura 2, que em ambas as classificações os Pokémons sem tipo secundários são a classificação majoritária, seguido pelo tipo Flying (voador), que se destaca para os Pokémons lendários. Para os okémons lendários também se salientas os tipos dragão, Fighting (lutador) e psíquico, enquanto que para os Pokémons não lendários temos os tipos Poison (venenoso), Ground (terrestre) e psíquico no top 5 de mais recorrentes.

2.0.3 Lançamento por gerações

Nessa subseção iremos analisar a quantidade de Pokémons em cada geração e comparar com o desenvolvimento de novos Pokémons lendários, começaremos analisando a frequência de Pokémons por geração na Figura 3.

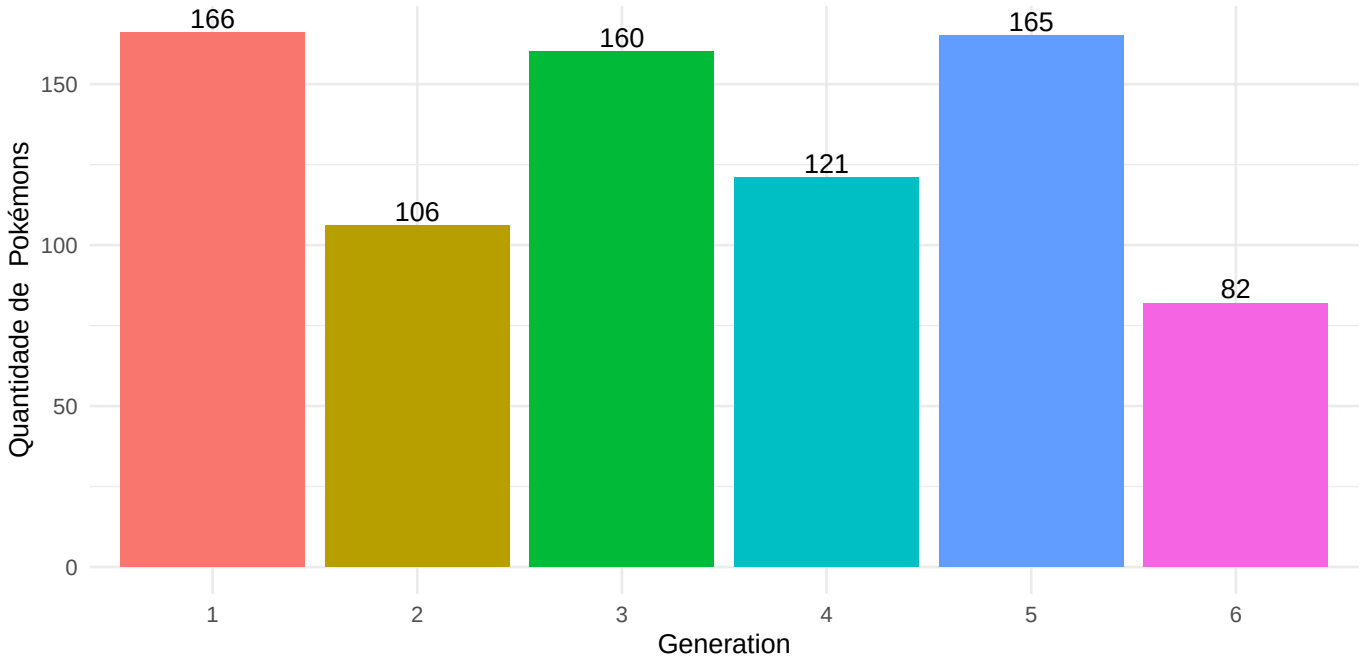

Figura 3: Frequência lançados por geração

Com a Figura 3, podemos notar uma "sazonalidade" para o desenvolvimento de Pokémons, visto que as gerações ímpares se sobresãem em número com todas com pelo menos 160 Pokémons novos e para as gerações pares tem-se uma queda notável, com um mínimo 82 novos Pokémons na geração 6 e um máximo de 121 na geração 5. A primeira geração ainda é imbatível com a quantidade de lançamentos de Pokémons, 166, no entanto, ganha por apenas 1 da quinta geração. Para o desenvolvimento de Pokémons lendários podemos ver a representação com a Figura 4.

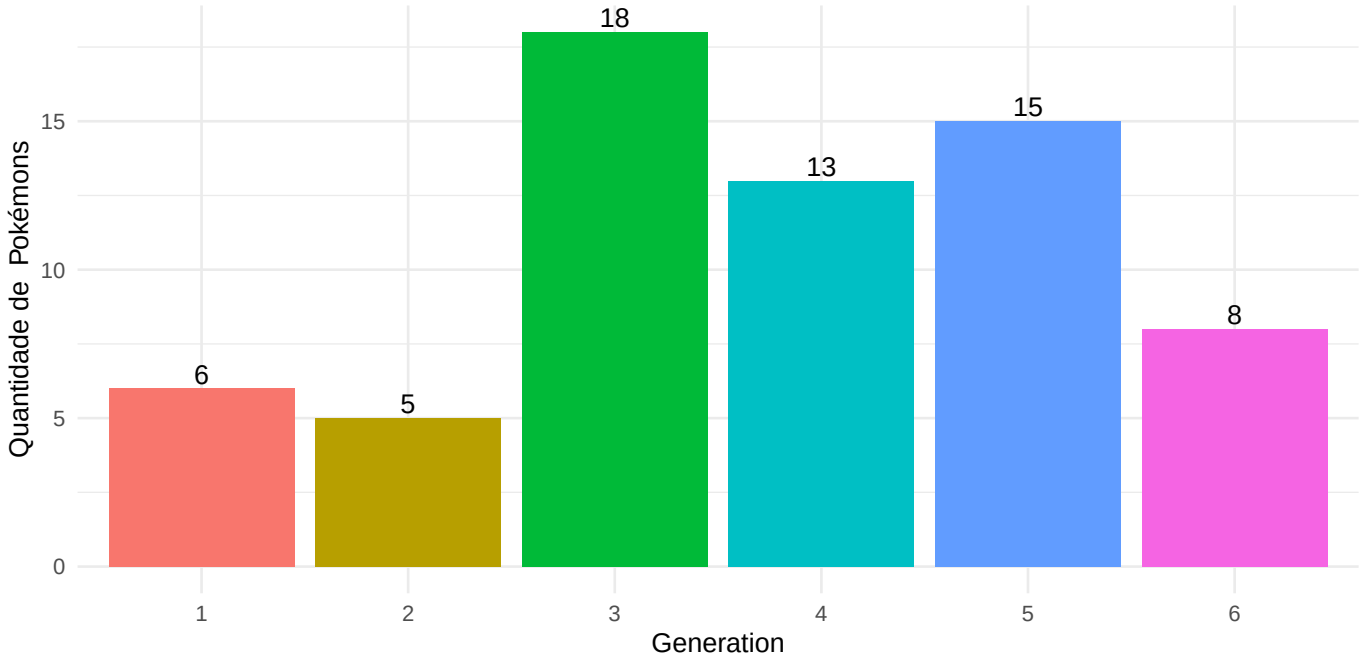

Figura 4: Frequência de novos Pokémons lendários por geração

Antagônico a Figura 3, na Figura 4 a terceira e quinta gerações se evidenciam, com a criação de Pokémons lendários, o fato deve-se que nas primeiras gerações os lendários estavam em torno do "Trio de Aves Lendárias" e "Trio de Cães Lendários", enquanto que nas próxima gerações foi se desenvolvendo e explicando a criação e manutenção do Universo Pokémon por suas espécies lendárias.

3 Análise Preditiva

Vamos aqui utilizar como variável de desfecho a classificação do Pokémon, sendo portanto Lendário ou não, vamos testar a saber 3 modelos para predição: Random Forest, Regressão logística e XGBoost. Como métrica de escolha de modelo vamos utilizar a sensibilidade e não a acurácia como erroneamente muitos fazem, pois como vimos, a proporção de Pokémons não lendários é 0.08, portanto se os modelos predizerem não lendário para todas observações teremos 92% de acurácia.

3.1 Regressão Logística

Regressão logistíca é um dos principais modelo estatístico atuais, pode ser descrito [1] como um modelo linear generalizado (MLG). Vamos considerar p a probabilidade de sucesso de uma certa variável binária, ou seja uma variável que tem distribuição Bernoulli.

O MLG usando como função de ligação logit pode ser escrito da seguinte maneira:

$$\log\left(\frac{p}{1-p}\right) = \sum_{i=1}^{n} \beta_i X_i \quad \text{onde} \quad X_0 = 1$$

Definindo $\sum_{i=1}^{n} \beta_i X_i$ como η fica fácil ver que p pode ser escrito como:

$$p = \frac{e^{\eta}}{1+e^{\eta}}$$

Apesar do modelo de regressão logística ser mais utilizado em análise inferencial, podemos também fazer predições de classes binárias se colocarmos um limiar para a saída (p) do modelo ser classificado como de certa classe, em outras palavras se $p \geq T$, onde T é um certo limite pré estabelecido, como não temos em mãos o modelo populacional trabalhamos com a predição $\hat{p}$ para a classificação do Pokémon ser lendário, aqui utilizamos $T = 0.5$.

3.2 Random Forest

Árvores de decisão são modelos que já existem a um certo tempo, apesar de terem uma grande vantagem em interpretabilidade são fracas em termos preditivos, assim a idéia de Random Forest é combinar diversas árvores alterando (bootstrap) o conjunto de treinamento de cada uma elas para gerar diversidade na predição, as árvores podem individualmente não serem fortes preditoras mas queremos no geral a predição combinada delas seja. Uma peculiaridade da Random Forest é que podemos ver a importância[2] das variáveis, o que é ilustrado pela Figura 5.

[1] Ou também um caso simples de uma neural network.

[2] Importância aqui: Decréscimo médio na impureza.

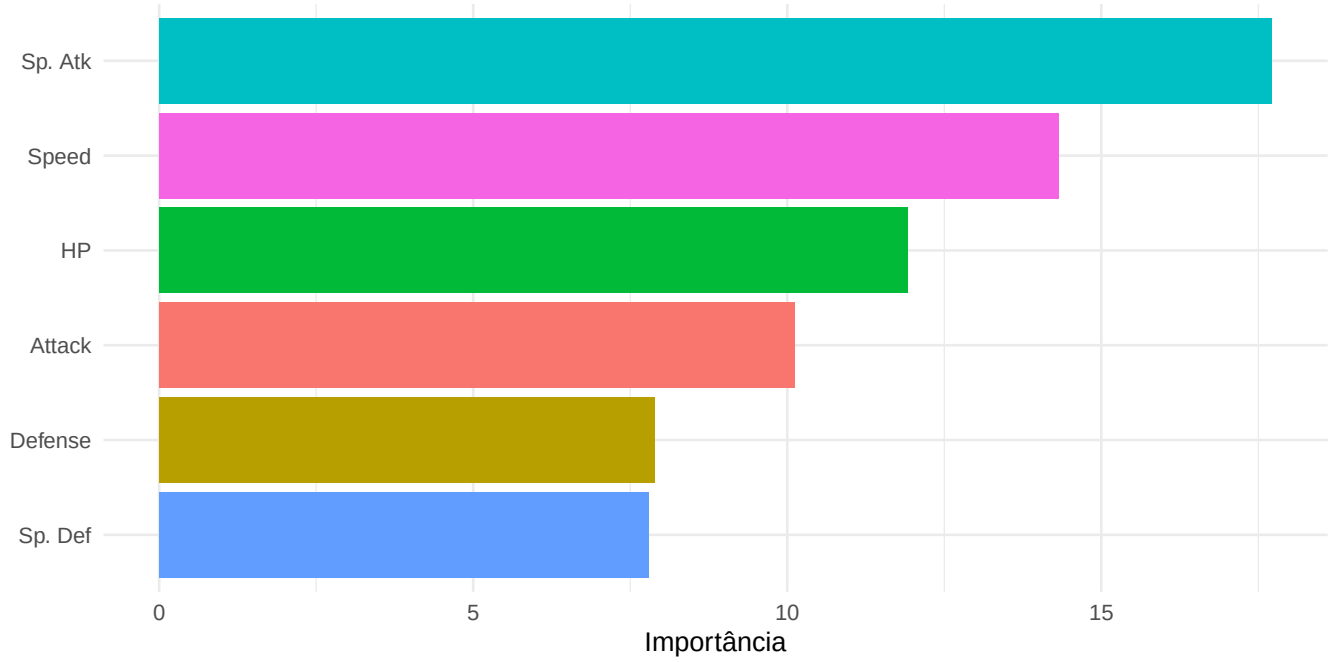

Figura 5: Importância das variáveis na Random Forest nos dados de treino

3.3 XGBoost

É notável que algoritmos de boosting atualmente são o estado da arte para dados estruturados, nele as árvores vão crescendo usando informações das árvores anteriores, isso quer dizer que não fazemos bootstrap dos dados igual em Random Forest, mas cada árvore trabalha com uma versão diferente dos dados originais, vamos aqui ajustar XGBoost para comparar com os modelos anteriores.

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 3: 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

Pode-se ver que em regressão logistica, que de 10 predições para lendário 9 foram corretas, porém para os 18 casos que eram lendários somente 50% foi corretamente classificado.

Tabela 4: Matriz de confusão para os dados de teste no modelo de Random Forest

	Predição		
Legendary	1	0	Total
1	9	9	18
0	2	180	182
Total	11	189	200

Em Random Forest nota-se que de 11 predições para lendário 9 foram corretas, porém assim como para a regressão logística 50% dos Pokémon lendários foram incorretamente classicados.

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

	Predição		
Legendary	1	0	Total
1	10	8	18
0	2	180	182
Total	12	188	200

Ao contrário dos modelos anteriores que tiveram 9 predições corretas para lendário, o XGBoost teve 10.
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

Temos como interesse principal o `ppv` e `sens`, pois estamos preocupados em predizer se o Pokémon **é lendário**, portanto as medidas que focam em acertos de não lendários não são de tamanha importância.
A Figura ref fornece o vislumbre de como as métricas se comportam para os 3 modelos ajustados na validação cruzada

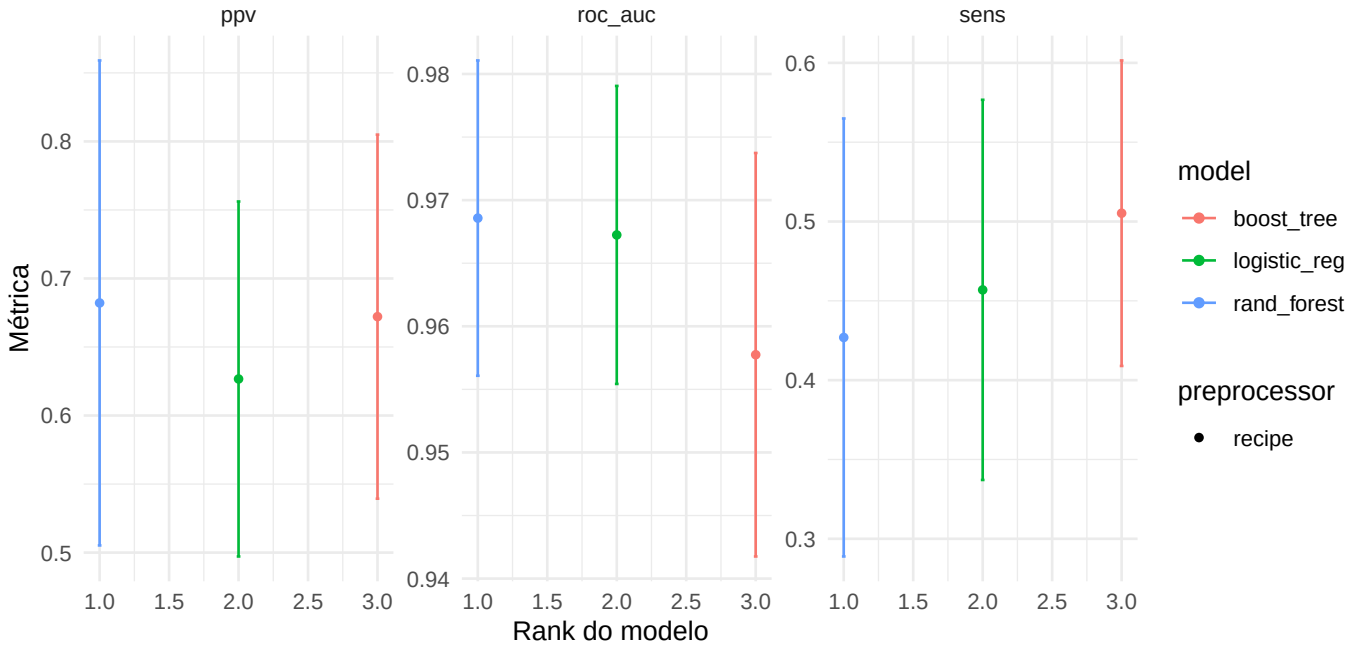

Figura 6: Métricas dos modelos na validação cruzada

E para os dados de teste temos:

Tabela 6: Métricas nos dados de teste

Modelo	sens	spec	ppv	npv	roc_auc
Regressão Logistica	0.500	0.995	0.900	0.953	0.984
Random Forest	0.556	0.989	0.833	0.957	0.988
XGBoost	0.556	0.989	0.833	0.957	0.981

Assim vemos pela tabela 6 que o modelo que maximizou a sensibilidade é c("Random Forest", "XGBoost")

4 Análise Inferencial

Nessa seção verificaremos uma etapa importante na análise de um ajuste de um modelo de regressão, a análise inferencial, em que busca encontrar possíveis distorções das suposições do modelo, principalmente deturpações no componente aleatório e observações discrepantes, juntamente com análise de resíduos, adequação da distribuição proposta, assim então validando o modelo. Ademais a análise inferencial também se ocupará de testes de hipóteses, apresentação, seleção do modelo e verificação das variáveis que mais influenciam a predição sendo negativa ou positivamente, determinando se a perspectiva tem sentido prático.

4.1 Análise de Dignóstico

Percebe-se que a análise e detecção de pontos influentes é um tópico relevante para a avaliação e validação de um modelo, assim nessa subseção iremos avaliar a existência de observações atípicas, isto é, pontos que exercem peso desproporcional nas estimativas dos parâmetros do modelo de Regressão Logística, por isso sucederemos com análise de pontos de avalanca, distância de cook, dffits e envelope simulado.

4.1.1 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, então na Figura 7, temos os valores da medida de alavancagem para cada observação.

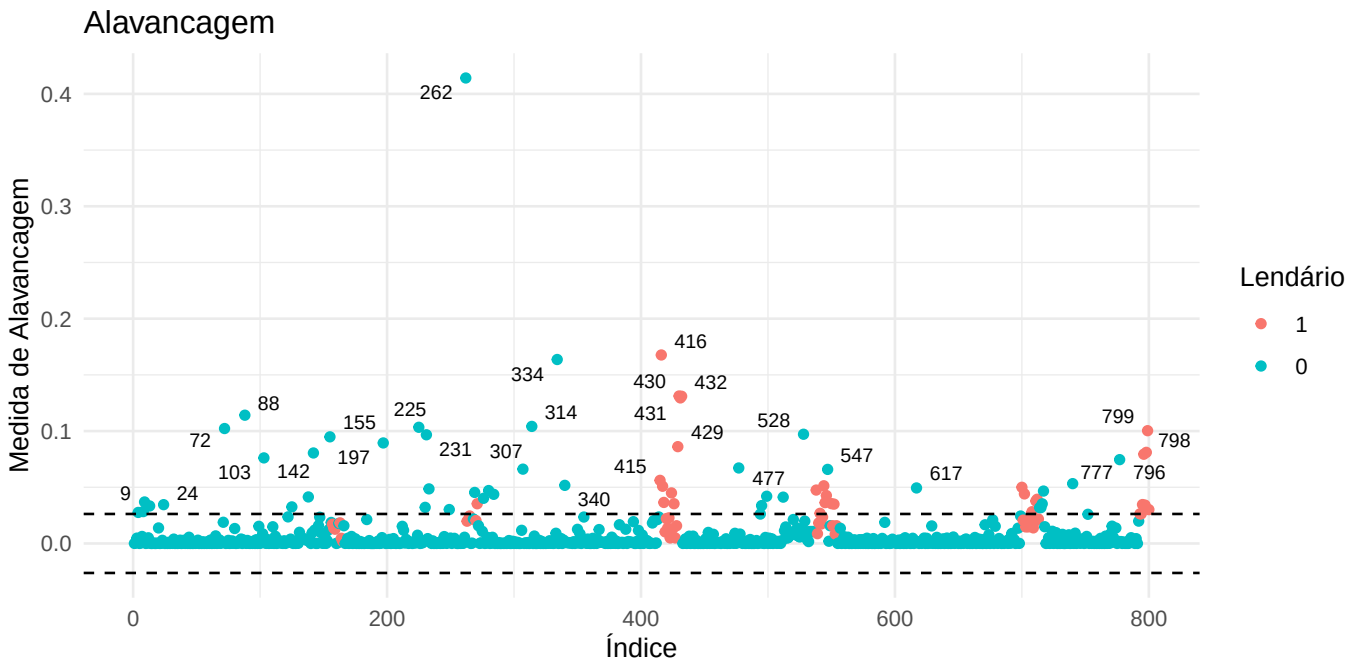

Figura 7: Medidas de Alavancagem para o Modelo 1

Com a Figura 7, tem-se que a observação que possui a maior alavancagem é o ponto 262 (o PokémonBlissey), devido ao fato que este Pokémon é o que mais achata a Figura, mas também temos outras observações acima do limite da medida de alavancagem estipulado.

4.1.2 Distância de Cook

Outra medida interessante para ponto aberrantes é a Distância de Cook, que mede essencialmente a influência das observações sobre os parâmetros e o ajuste, avaliando a influência de o que pequenas pertubarções nas variâncias das observações causam nas estimativas dos parâmetros. Ou de forma simplificada, temos a influência da observação i sobre todos os n valores ajustados,

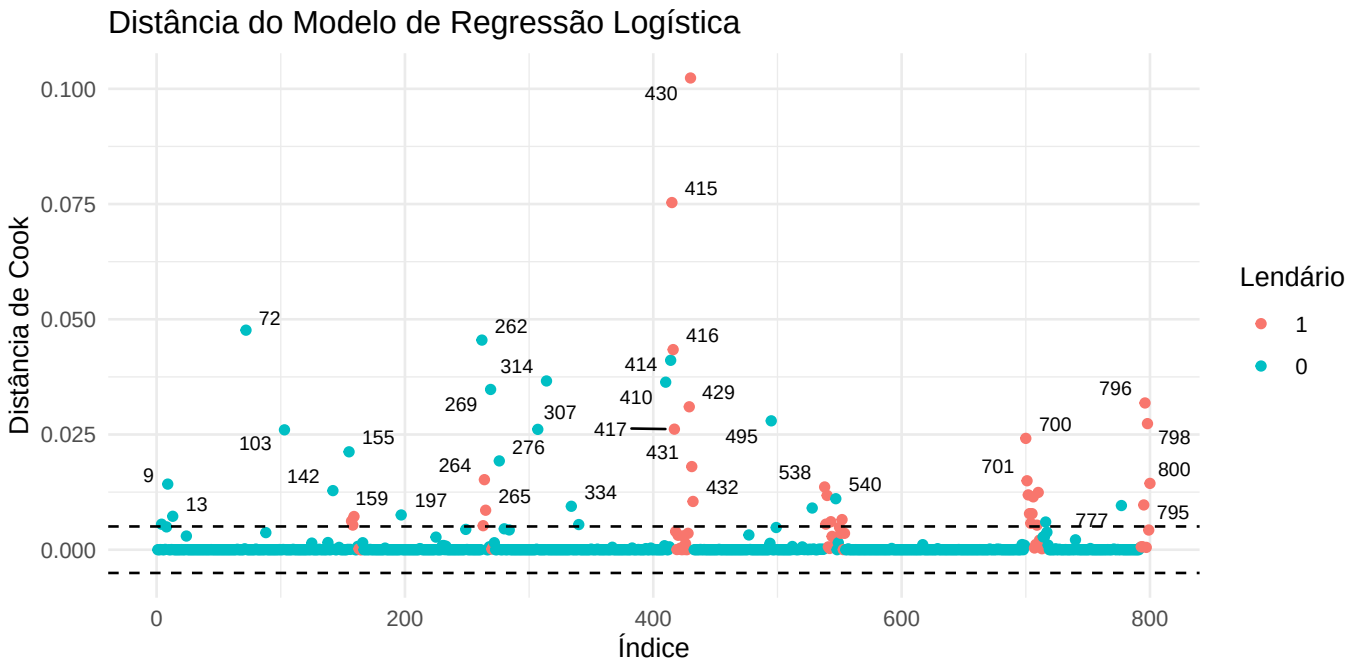

Figura 8: Distância de Cook para o Modelo 1

Com a Figura 8, tem-se que as observações que possuem maior valores da distância de cook são: 72, 262, 414, 415, 416 e 430 (sendo os Pokémons conhecidos como: Mega Alakazam, Blissey, Mega Metagross, Regirock, Regice, Deoxys Attack, respectivamente), devido ao fato que estes Pokémons são os que estão mais distantes do intevalo definido, no entanto também existem outras observações fora dos limites.

4.1.3 DFFITS

A medida DFFITS pode ser considerada uma medida complementar ou concorrente a distância de Cook, tendo o próposito de medir a influência das observações nos parâmetros de posição e escala (mas não simultaneamente), assim informam o grau de influência que a observação i tem sobre o valor seu próprio valor ajustado $\hat{y}_i$, tem-se que:

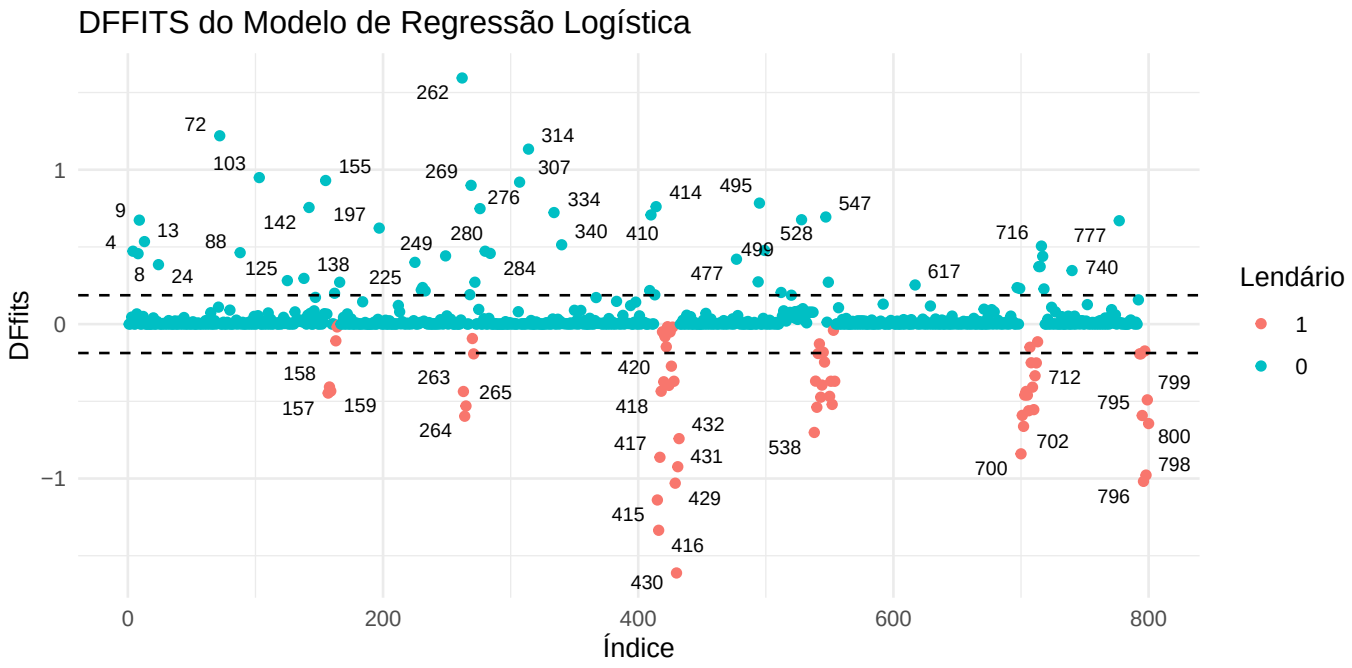

Figura 9: DFFITS do Modelo 1

Com a Figura 9 percebe-se que as observações que podem ser consideradas pontos influentes de acordo com o DFFITS são: 262, 415, 416, 429 e 430, (sendo os Pokémons conhecidos como: Blissey, Regirock, Regice, Deoxys, Deoxys Attack), visto que são os indíviduos que mais distorcem a Figura 9.

4.1.4 Resíduos

A análise específica 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, então utilizando o resíduo deviance as suas ocorrências podem ser observadas na Figura 10.

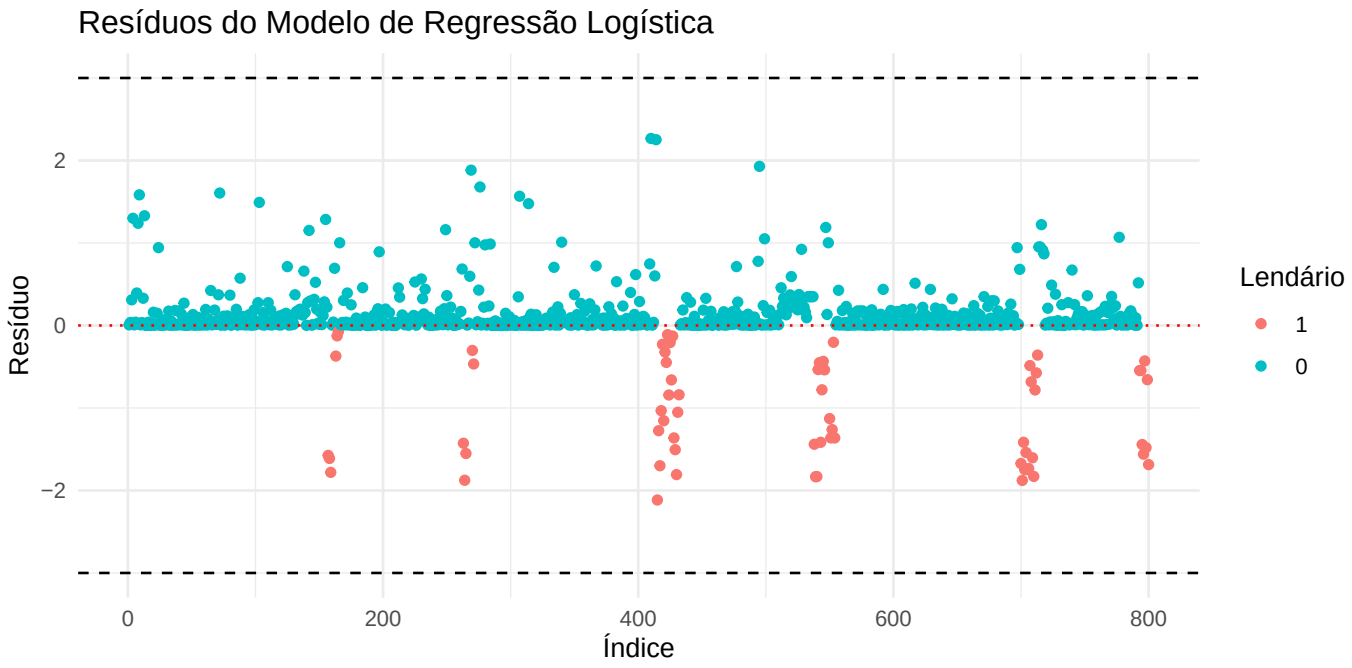

Figura 10: Resíduos Deviance do Modelo 1

Com a Figura 10 verificamos que todos as ocorrências estão dentro dos limites.

4.1.5 Envelope Simulado

Finalizamos a análise de influência com o envolope simulado que permite uma melhor comparação entre os resíduos e os percentis da distribuição, fornecendo um vislumbre se a distribuição é adequada para o ajuste, como percebemos na Figura 11.

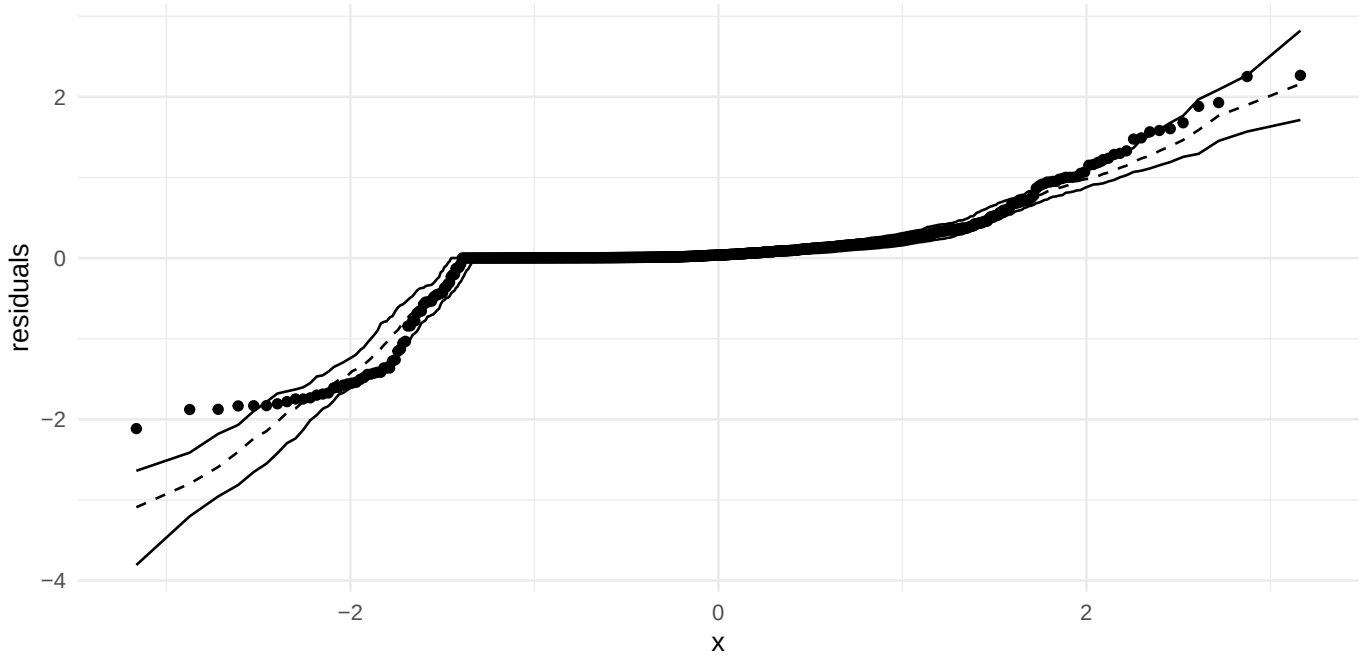

Figura 11: Envelope Simulado dos dados com a Distribuição Binomial

Analisando a Figura 11, pode-se identificar que existem alguns pontos fora das bandas simuladas, fornecendo indícios de que existem pontos influentes ou que talvez a distribuição não é adequada para o ajuste.

4.1.6 Decisão

Tendo em consideração toda a análise de influência e o objetivo da modelagem, decidimos por não remover nenhuma observação possivelmente influente, dado que o modelo está sendo usado para avaliar e prever condições raras na população (Pokémons lendários) e essa condição leva a valores extremos das variáveis preditoras, pois estes Pokémons possuem atributos maiores devidos a sua importância no universo fictício, desde proteger certas regiões, criar tempestades, gelo ou até mesmo o universo, devido a esses fatos eles possuem atributos de maior grau. Dado que, os possíveis pontos influentes em sua maioria são dados por Pokémons lendários e com o decorrer do crescimento do universo Pokémon se desejassemos predizer Pokémons lendários com o modelo proposto nesse trabalho e os status parecidos com os dos pontos aberrantes, o modelo não estaria adequado para essa situação. No entanto, buscamos remover as observações discrepantes até que pudessemos chegar em condições melhores das análises de influência, que podem ser verificadas aqui.

4.2 Testes

Agora devemos prosseguir para a avaliação do modelo de Regressão Logística ajustado, começando com a análise de significância das variáveis dadas pelo seguinte teste de hipótese:

$$H_0 : \beta_i = 0 \text{ (Covariável não-significativa)}$$

$$H_0 : \beta_i \neq 0 \text{ (Covariável significante)}$$

O qual podemos analisar com a Tabela 7.

Tabela 7: Estatísticas do Modelo 1

term	estimate	std.error	statistic	p.value
(Intercept)	21.877	2.443	8.96	<0.001
HP	-0.034	0.009	-3.92	<0.001
Attack	-0.018	0.007	-2.69	0.007
Defense	-0.033	0.008	-3.98	<0.001
‘Sp. Atk‘	-0.036	0.007	-5.01	<0.001
‘Sp. Def‘	-0.043	0.009	-4.77	<0.001
Speed	-0.050	0.010	-5.26	<0.001

A Tabela 7 detalha, algumas estatísticas muito importantes sobre as covariáveis, mas principalmente informa que todas as variáveis preditivas são significativas.

4.2.1 Critério de seleção de modelos

Existem vários procedimentos para a seleção de modelos, no entanto, nos utilizaremos principalmente de dois critérios AIC e BIC, que são processos de minimização que não envolvem testes, sendo a ideia básicia buscar um modelo que seja parcimonioso, ou seja, bem ajustado e com um número reduzido de parâmetros, a diferença entre os critérios são suas formas para penalização. Desse modo, a Tabela 8, tem vários critérios de seleção:

Tabela 8: Critérios de Seleção do Modelo para a Regressão Logistica

null.deviance	df.null	logLik	AIC	BIC	deviance	df.residual	nobs
451	799	-88.4	191	224	177	793	800

Valendo-se da Tabela 8, temos que o critério de AIC, 190.889, e BIC, 223.682, não contém valores elevados, para efeito de comparação se levarmos em conta os modelos com remoção de pontos influentes, o AIC do modelo proposto está bem próximo do modelo com poucas remoções e não está a uma diferença considerável do modelo com muitas remoções.

4.2.2 Modelo Final e Inferências

Nessa subseção iremos apresentar a estrutura do modelo, assim como a sua interpretação, avaliando a influencia das variáveis dependentes e verificando se converge com a realidade dos Pokémons lendários.

Assim temos que a estrutura final do modelo ajustado, com y sendo a classificação como Pokémon lendário, é:

$$\log\left[\frac{P(\mathrm{y}=1)}{1-P(\mathrm{y}=1)}\right] = \beta_0 + \beta_1(\mathrm{HP}) + \beta_2(\mathrm{Attack}) + \beta_3(\mathrm{Defense}) + \beta_4(\mathrm{Sp.\ Atk}) + \beta_5(\mathrm{Sp.\ Def}) + \beta_6(\mathrm{Speed})$$

No entanto, podemos realizar algumas manipulações algébricas, para buscar a influência de cada cováriavel na chance dos Pokémons serem preditos como lendários, começaremos com a aplicação do exponencial (e):

$$\frac{P(\mathrm{p}=1)}{1-P(\mathrm{p}=1)} = e^{\beta_0+\beta_1(\mathrm{HP})+\beta_2(\mathrm{Attack})+\beta_3(\mathrm{Defense})} * e^{\beta_4(\mathrm{Sp.\ Atk})+\beta_5(\mathrm{Sp.\ Def})+\beta_6(\mathrm{Speed})}$$

Com isso, já conseguimos verificar a influência de cada variável na chance de ser lendário, considerando todos os outros atributos constantes:

- **HP:** A adição de 1 de HP, acresce em 0.966, a chance de um Pokémon ser categorizado como lendário;
- **Attack:** O acréscimo de uma unidade de Attack, aumenta em 0.982, a chance de um Pokémon ser predito como lendário;

- **Defense:** A soma de 1 de Defense, acrescenta em 0.968, a chance de um Pokémon ser definido como lendário;
- **Sp. Attack:** A inclusão de 1 de Sp. Attack, amplifica em 0.965, a chance de um Pokémon ser delineado como lendário;
- **Sp. Defense:** O aumento de 1 de Sp. Defense, adiciona em 0.958, a chance de um Pokémon ser categorizado como lendário;
- **Speed:** A amplicação de 1 de Speed, incorpora em 0.951, a chance de um Pokémon ser predito como lendário;

Consoante com as Tabelas 2 e 1, percebemos que a influência dos preditores tem fundamento concreto, em virtude das similaridades das médias e medianas dos atributos, ficando em níveis semelhantes e assim como na Tabela 1, o atributo que se destaca é principalmente o Attack e tem-se uma similaridade com a influência das covariáveis no modelo de Random Forest, dado pela Figura 5, onde o status de Sp. Attack é que se sobresãe.

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.