Pokémon Prediction

Predicting legendary status of Pokémon

Marc Secasan
STA5737
MMXXV anno Domini

Predicting legendary vs non-legendary pokémon

Data from: https://www.kaggle.com/datasets/sarahtaha/1025-pokemon
Author: Sarah Taha, Data Science and Statistics Bachelor's Student

Legendary: A pokémon of great power, usually with domain over some natural phenomena. Often genderless, of single existence, and ancient.

Data

25 Variables - 1184 rows

Name : chr "bulbasaur" Entropy 10.20

National.Dex.. : int 1

Primary.Typing : chr "grass" Entropy 3.98

Secondary.Typing : chr "poison" Entropy 3.15

Secondary.Typing.Flag: chr "True" Entropy 0.99

Generation : chr "generation-i" Entropy 3.11

Legendary.Status : Factor w/ 2 levels "False"

Form : chr "Base" Entropy 1.13

Alt.Form.Flag : chr "False" Entropy 0.56

Evolution.Stage : Factor w/ 3 levels "1","2","3"

Number.of.Evolution : int 3

Height..in. : int 28

Weight..lbs. : int 15

Base.Stat.Total : int 318

Health : int 45

Attack : int 49

Defense : int 49

Special.Attack : int 65

Special.Defense : int 65

Speed : int 45

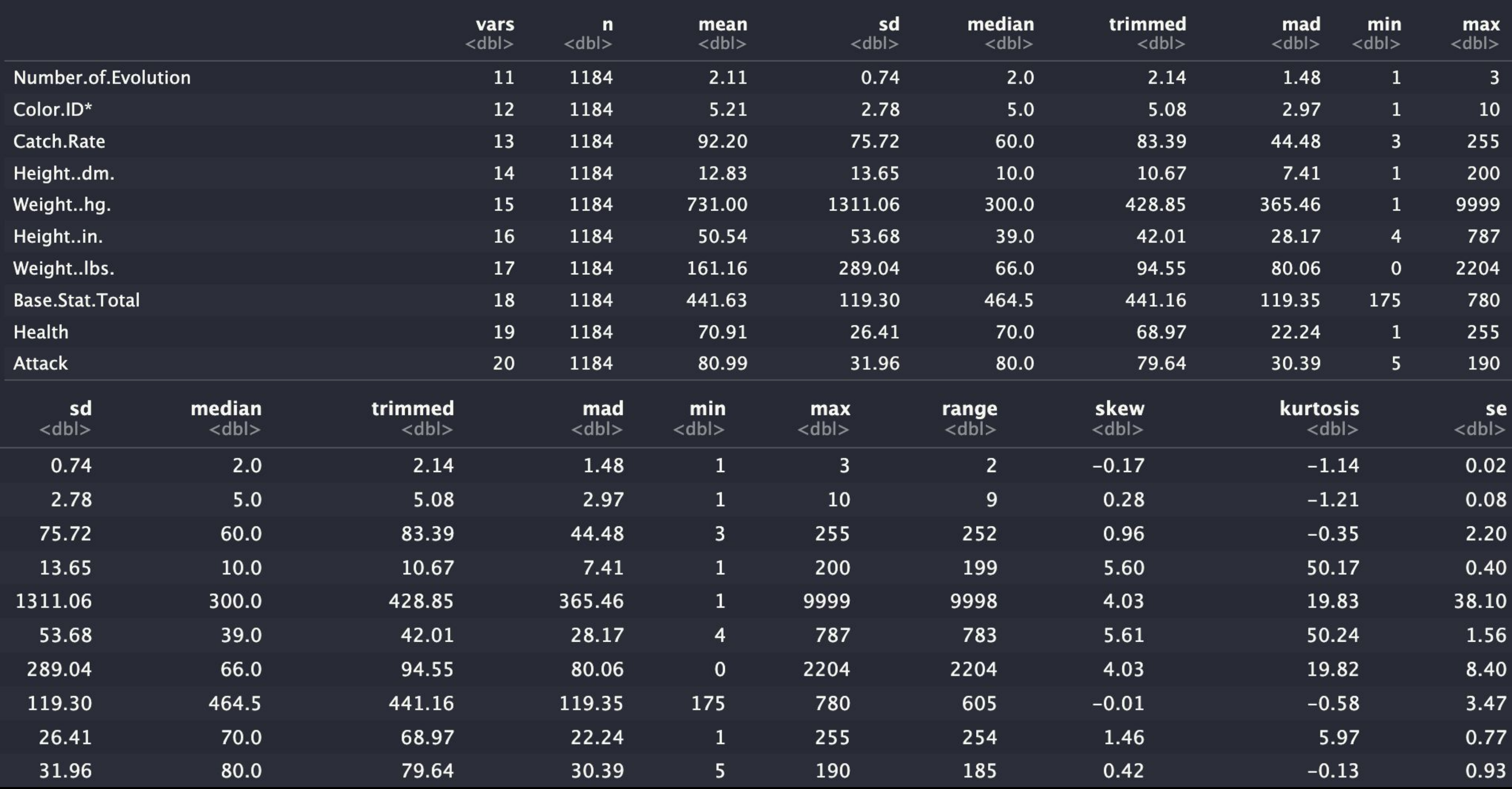
cluster : Factor w/ 3 levels "1","2","3"

Color.ID : chr "green"; Entropy 3.23

Catch.Rate : int 45

Height..dm. : int 7

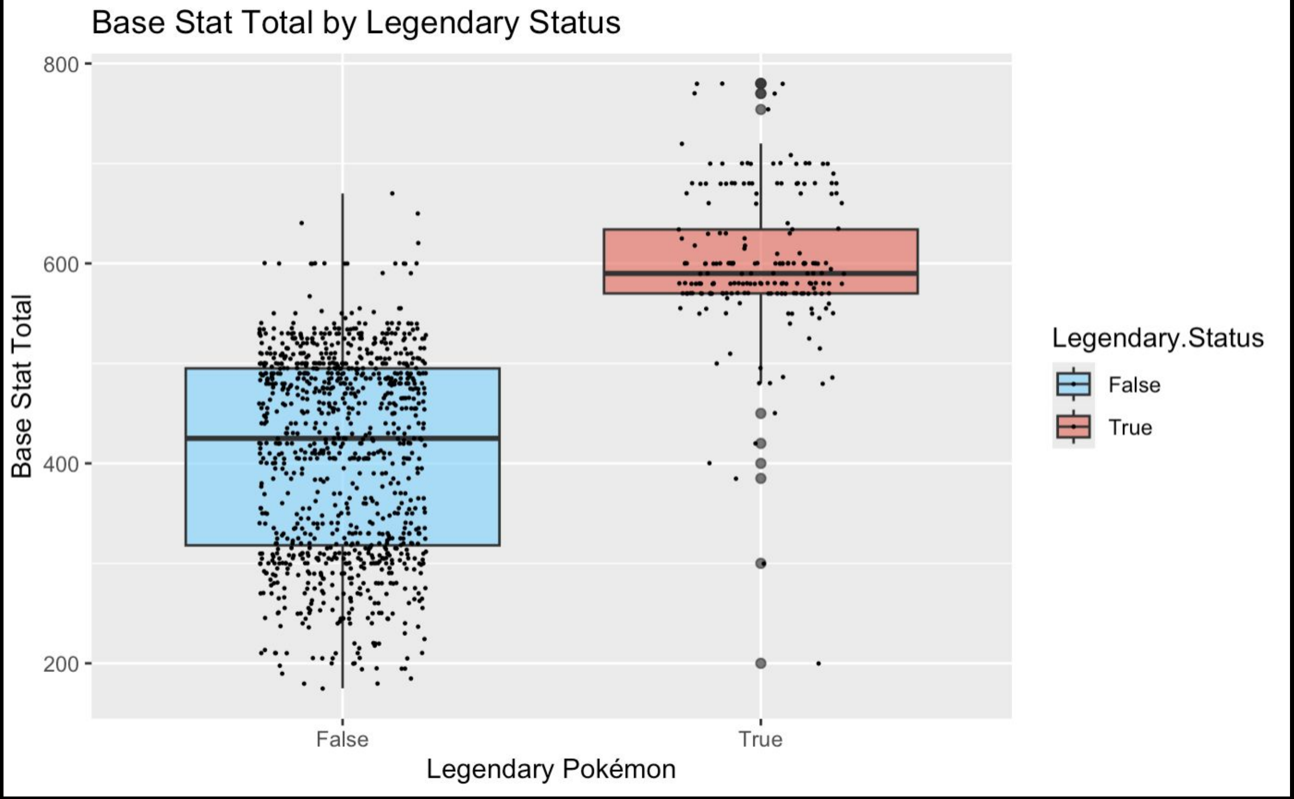
Weight..hg. : int 69



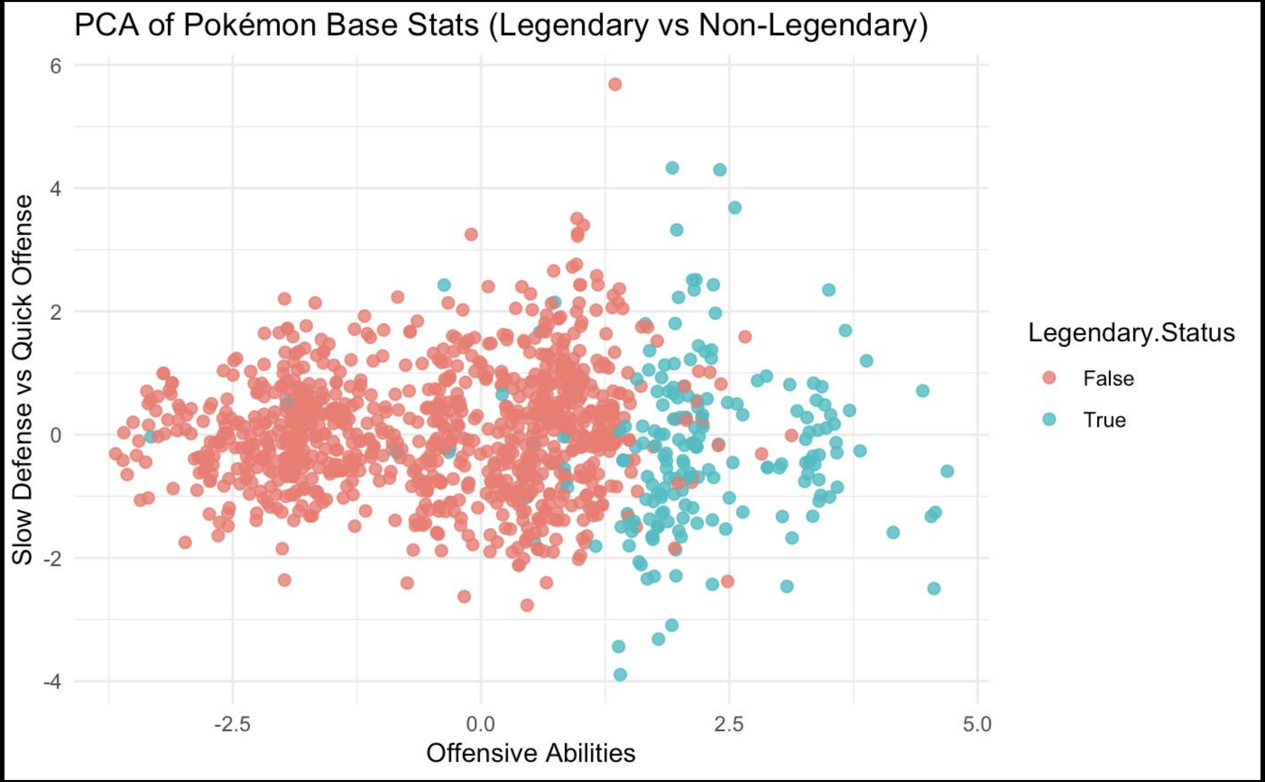
	vars <dbl>	n <dbl>	mean <dbl>	sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>
Number.of.Evolution	11	1184	2.11	0.74	2.0	2.14	1.48	1	3
Color.ID*	12	1184	5.21	2.78	5.0	5.08	2.97	1	10
Catch.Rate	13	1184	92.20	75.72	60.0	83.39	44.48	3	255
Height..dm.	14	1184	12.83	13.65	10.0	10.67	7.41	1	200
Weight..hg.	15	1184	731.00	1311.06	300.0	428.85	365.46	1	9999
Height..in.	16	1184	50.54	53.68	39.0	42.01	28.17	4	787
Weight..lbs.	17	1184	161.16	289.04	66.0	94.55	80.06	0	2204
Base.Stat.Total	18	1184	441.63	119.30	464.5	441.16	119.35	175	780
Health	19	1184	70.91	26.41	70.0	68.97	22.24	1	255
Attack	20	1184	80.99	31.96	80.0	79.64	30.39	5	190

sd <dbl>	median <dbl>	trimmed <dbl>	mad <dbl>	min <dbl>	max <dbl>	range <dbl>	skew <dbl>	kurtosis <dbl>	se <dbl>
0.74	2.0	2.14	1.48	1	3	2	-0.17	-1.14	0.02
2.78	5.0	5.08	2.97	1	10	9	0.28	-1.21	0.08
75.72	60.0	83.39	44.48	3	255	252	0.96	-0.35	2.20
13.65	10.0	10.67	7.41	1	200	199	5.60	50.17	0.40
1311.06	300.0	428.85	365.46	1	9999	9998	4.03	19.83	38.10
53.68	39.0	42.01	28.17	4	787	783	5.61	50.24	1.56
289.04	66.0	94.55	80.06	0	2204	2204	4.03	19.82	8.40
119.30	464.5	441.16	119.35	175	780	605	-0.01	-0.58	3.47
26.41	70.0	68.97	22.24	1	255	254	1.46	5.97	0.77
31.96	80.0	79.64	30.39	5	190	185	0.42	-0.13	0.93

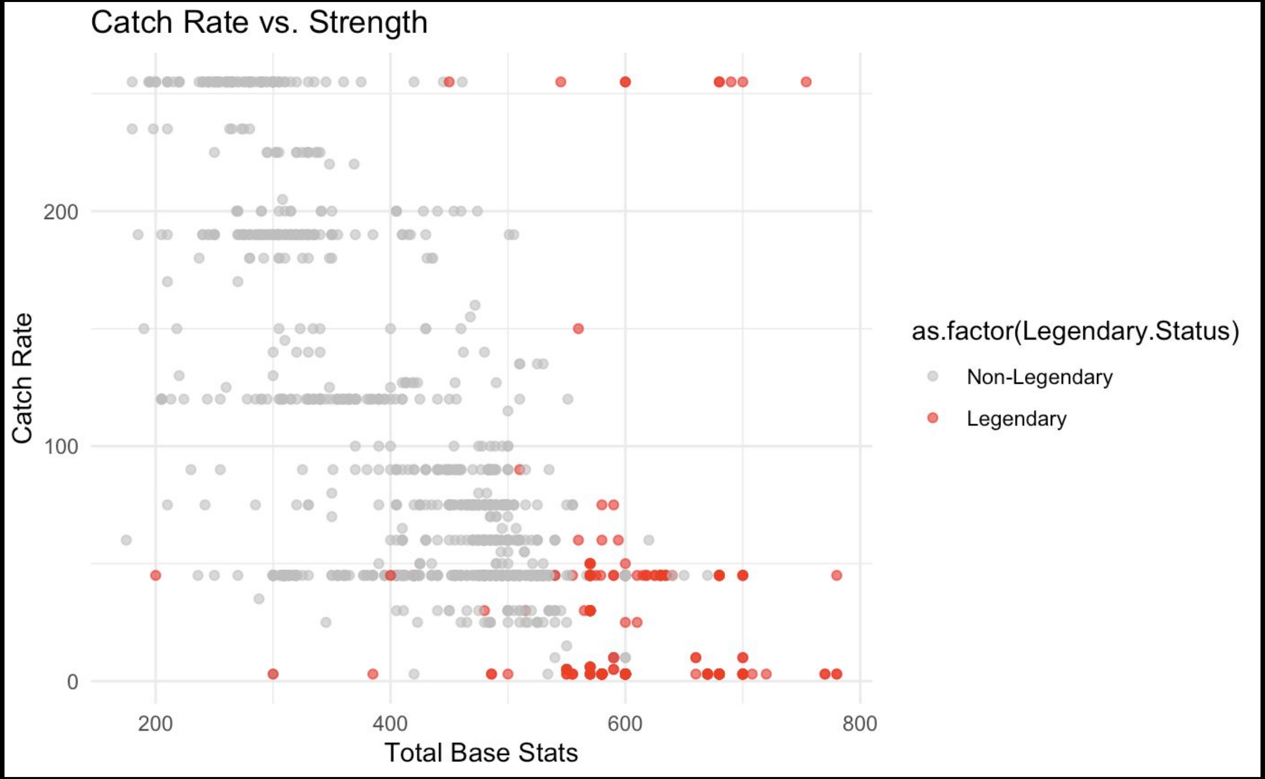
Data cont. - Visualization



Data cont. - Visualization

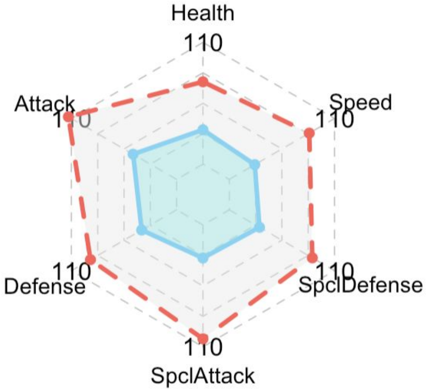


Data cont. - Visualization

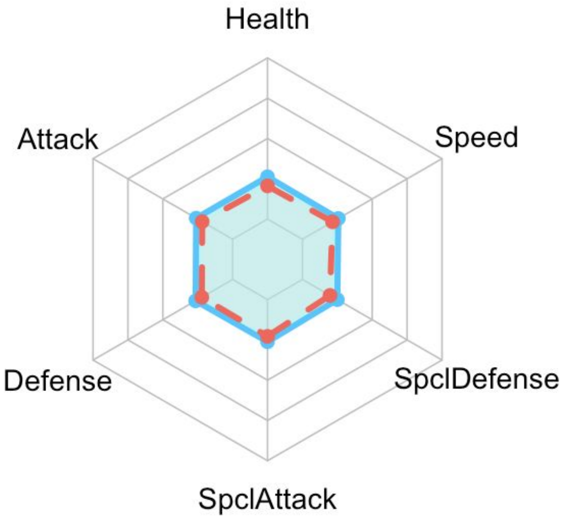


Data cont. - Visualization

Average Base Stats: Legendary vs Non-Legendary



Variance of Base Stats: Legendary vs Non-Legendary



Methodology & Results

Language: R
Packages: DBI, RSQLite, ggplot2, dplyr, Rtsne (t tests), fmsb (radar plots), psych (desc. stats), cvms (matrices), caret(log model), xgboost

SQLite used for database management system

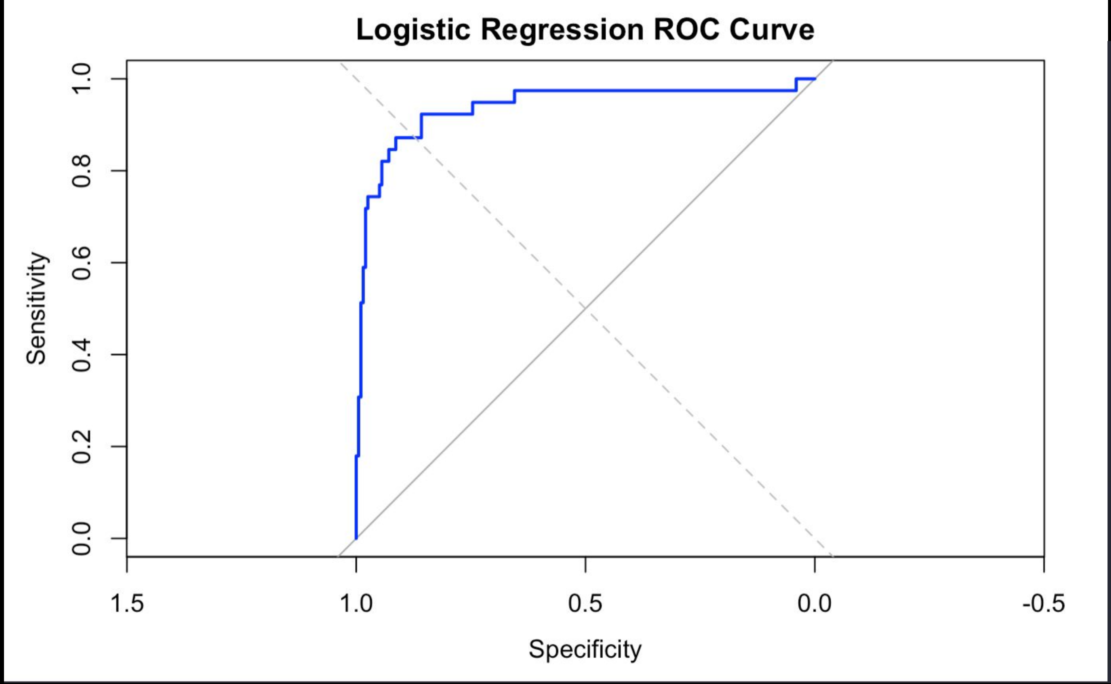
Methodology & Results

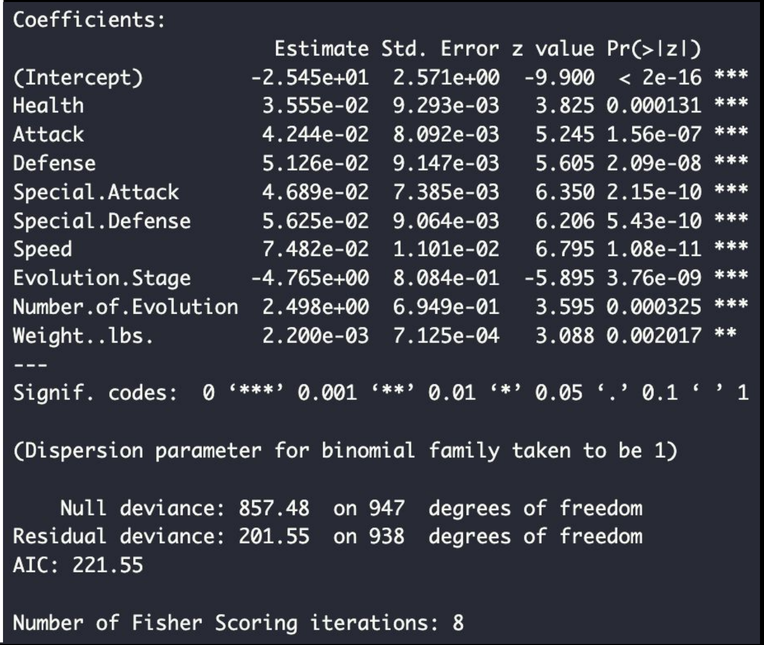
Logistic Model 1: log_model <- glm(Legendary.Status ~ ., data = train_data, family = "binomial")
AIC; 240.41
Fisher Scoring: 9
Number of predictors: 37 - (factor variables broken into separate columns)

Logistic Model 2: log_model_reduced <- step(log_model, direction = "both")
AIC; 221.55
Fisher Scoring: 8
Number of predictors: 9 - (all statistically significant)

Methodology & Results

Logistic cont. 92.8% accuracy

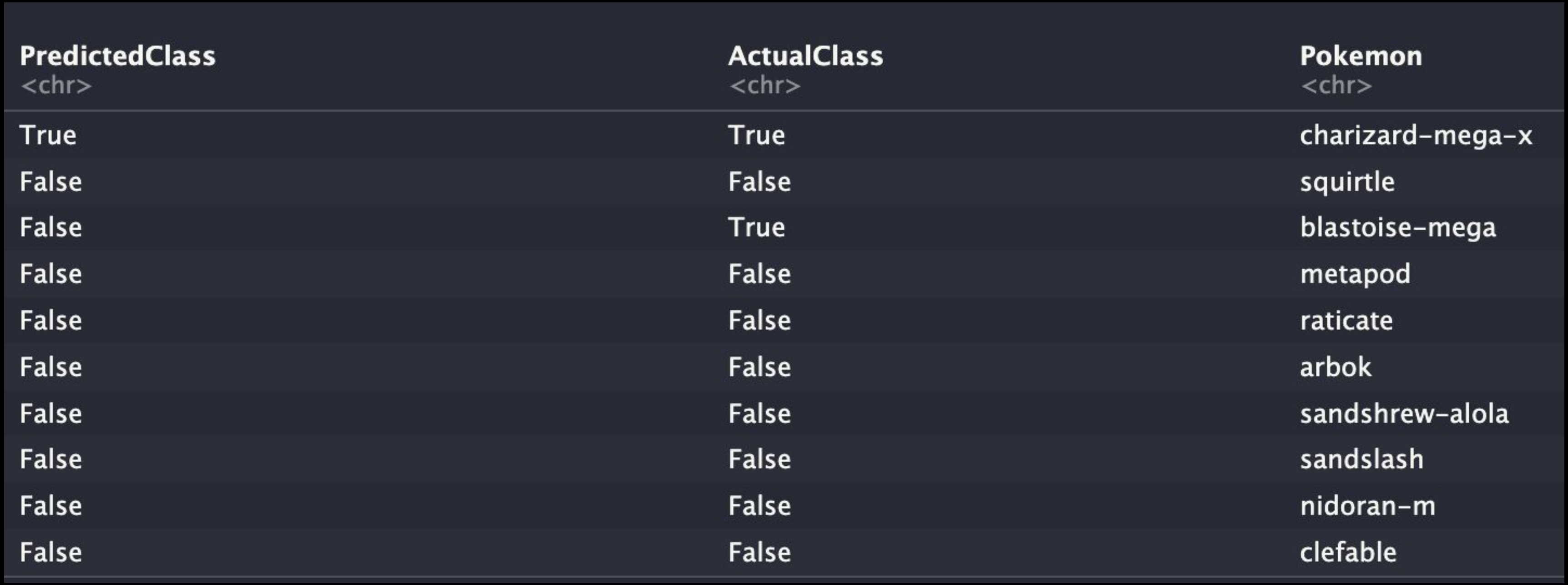




```
Coefficients:
                      Estimate Std. Error z value Pr(>|z|)    
(Intercept)         -2.545e+01  2.571e+00  -9.900  < 2e-16 ***
Health               3.555e-02  9.293e-03   3.825 0.000131 ***
Attack               4.244e-02  8.092e-03   5.245 1.56e-07 ***
Defense              5.126e-02  9.147e-03   5.605 2.09e-08 ***
Special.Attack       4.689e-02  7.385e-03   6.350 2.15e-10 ***
Special.Defense      5.625e-02  9.064e-03   6.206 5.43e-10 ***
Speed                7.482e-02  1.101e-02   6.795 1.08e-11 ***
Evolution.Stage     -4.765e+00  8.084e-01  -5.895 3.76e-09 ***
Number.of.Evolution  2.498e+00  6.949e-01   3.595 0.000325 ***
Weight..lbs.         2.200e-03  7.125e-04   3.088 0.002017 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 857.48  on 947  degrees of freedom
Residual deviance: 201.55  on 938  degrees of freedom
AIC: 221.55

Number of Fisher Scoring iterations: 8
```

Methodology & Results

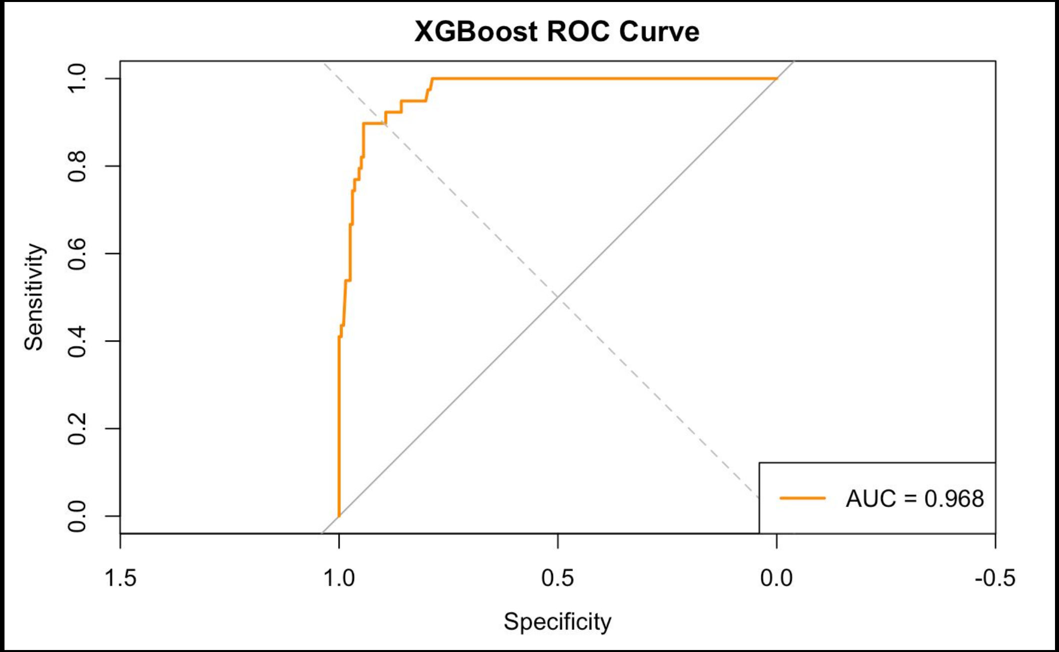
xgBoost 1: Accuracy: 93.8%
Variables of Importance: Catch Rate, special attack, attack, weight

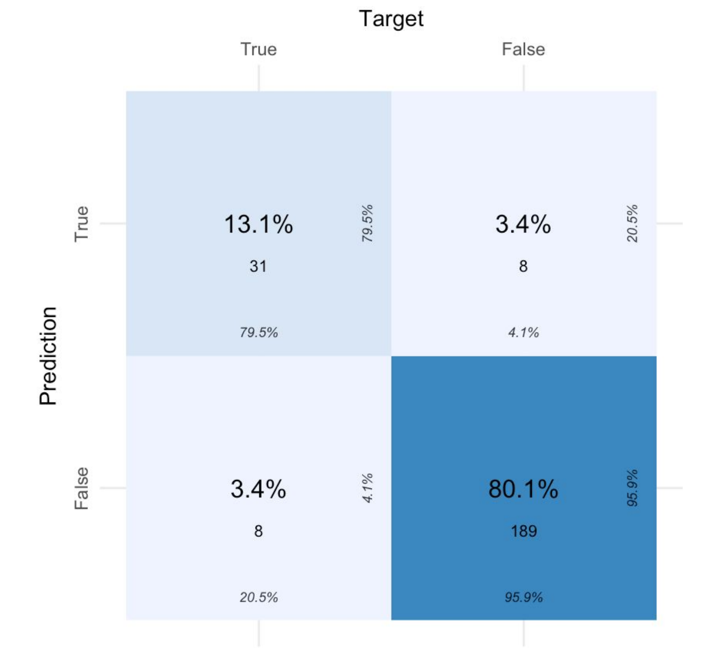


PredictedClass	ActualClass	Pokemon
<chr>	<chr>	<chr>
True	True	charizard-mega-x
False	False	squirtle
False	True	blastoise-mega
False	False	metapod
False	False	raticate
False	False	arbok
False	False	sandshrew-alola
False	False	sandslash
False	False	nidoran-m
False	False	clefable

Methodology & Results

xgBoost 1: Accuracy: 93.8%
Variables of Importance: Catch Rate, special attack, attack, weight

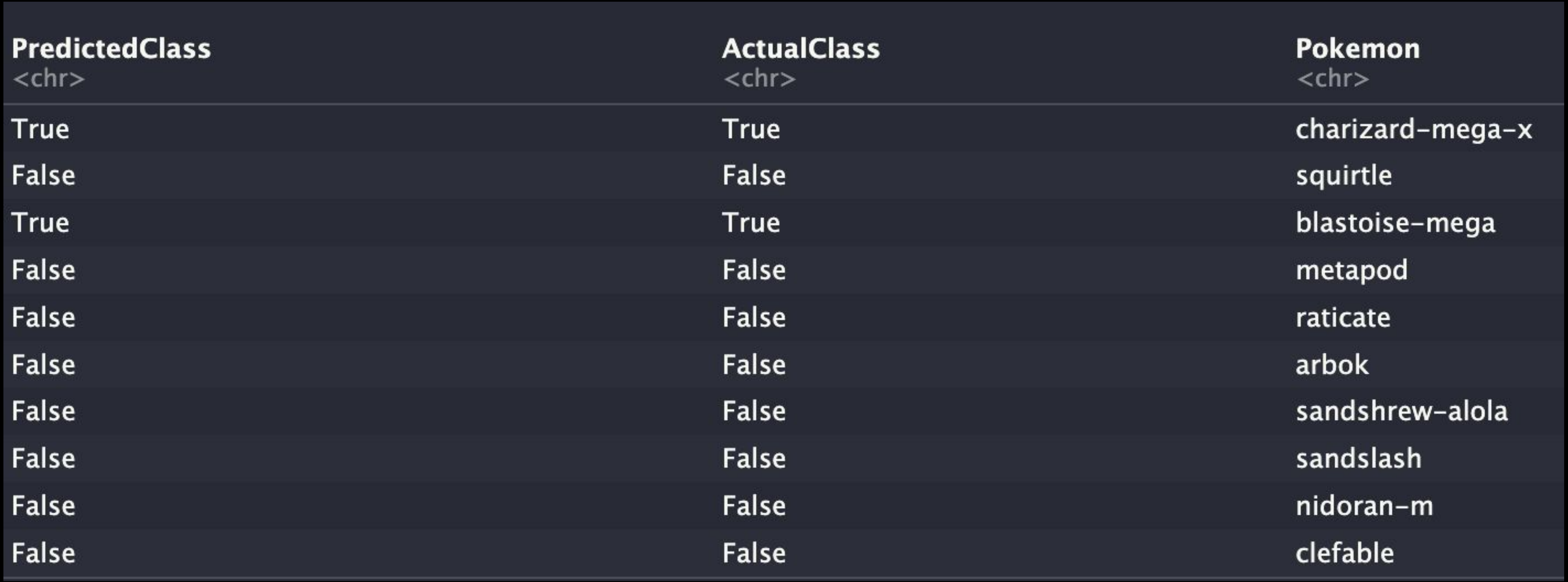




	Target	
Prediction	True	False
True	13.1% 31 79.5% 79.5%	3.4% 8 4.1% 20.5%
False	3.4% 8 20.5% 4.1%	80.1% 189 95.9% 95.9%

Methodology & Results

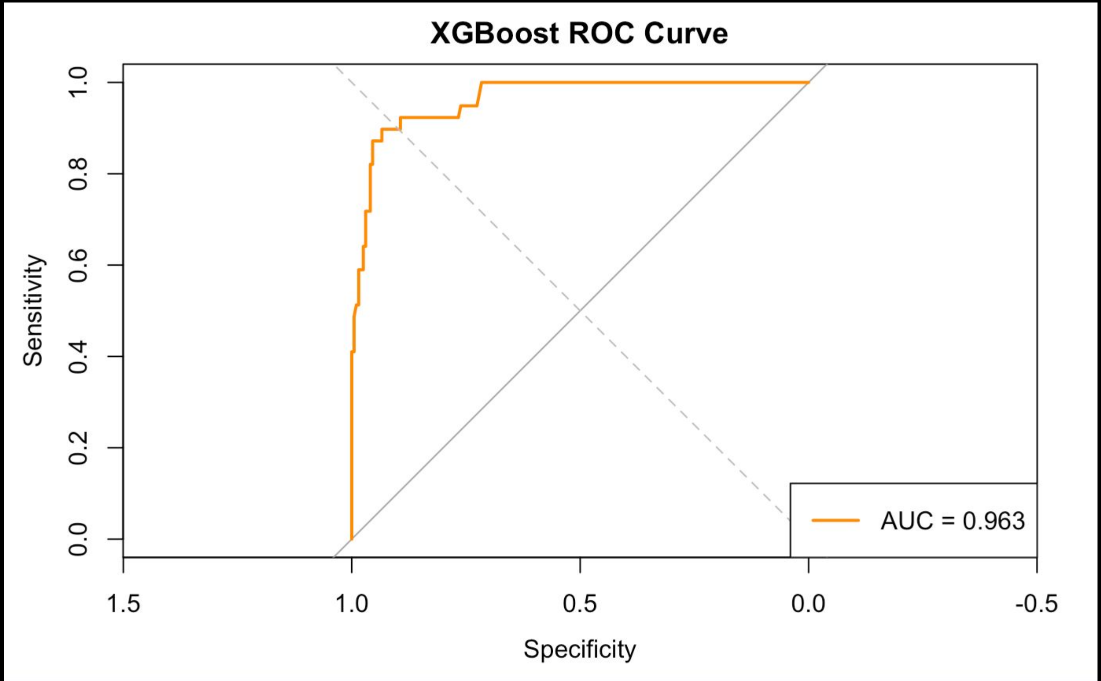
xgBoost 2: Accuracy: 92.4%
Variables: Catch Rate, special attack, attack, weight



PredictedClass	ActualClass	Pokemon
<chr>	<chr>	<chr>
True	True	charizard-mega-x
False	False	squirtle
True	True	blastoise-mega
False	False	metapod
False	False	raticate
False	False	arbok
False	False	sandshrew-alola
False	False	sandslash
False	False	nidoran-m
False	False	clefable

Methodology & Results

xgBoost 2: Accuracy: 92.4%
Variables: Catch Rate, special attack, attack, weight



Question:

How do Pokémon vary by generation