Final Project
Stakeholder Report

Statistics - 511

Carlos Madrid
Joel Borja
Xirui Huang

cmadrid21@apu.ed
jborja21@apu.edu
xiruihuang21@apu.edu
Department of Math and Physics
Azusa Pacific University
May 4, 2022

Intro/Background/Significance

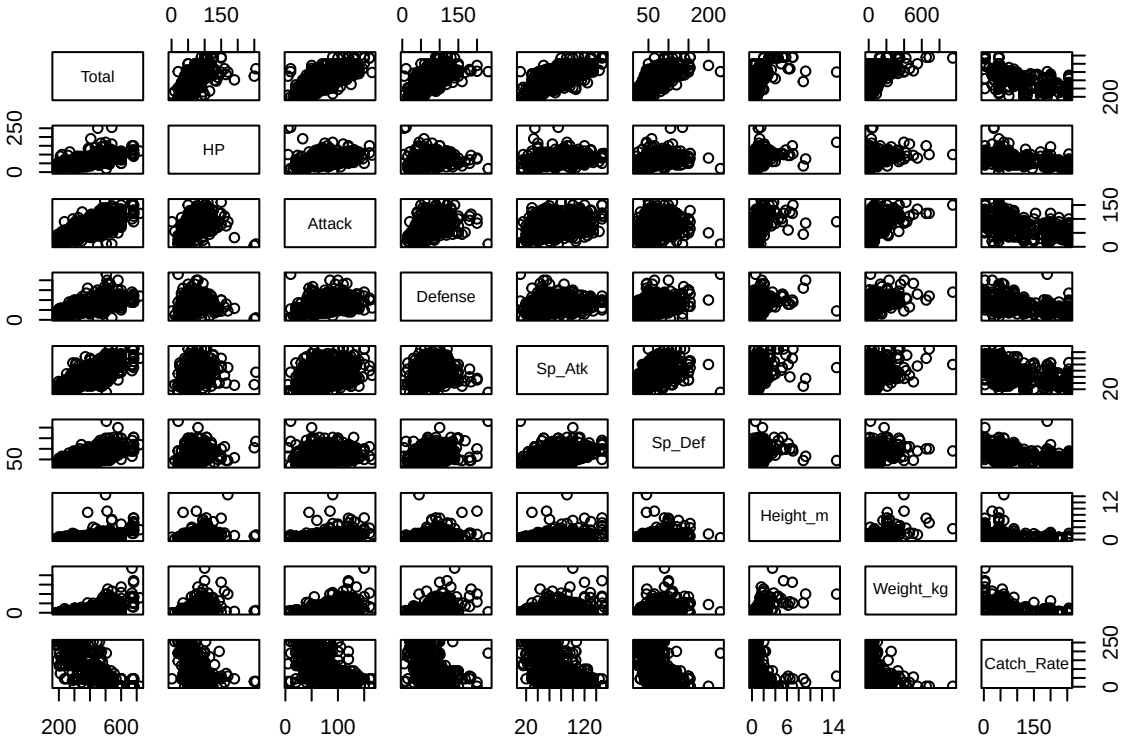
Pokémon is a media franchise that started in 1996. Pokémon are mythical animal-like monsters that live in the fictitious Pokémon World. Every new cycle of Pokémon brings in new generations. Pokémon are entered into competitions so that they may battle. The last one standing after various rounds is the winner. They fight by using different abilities. We examine the Pokémon's different abilities, physical attributes, attack power, defense, etc. Our goal is to identify what predicts a Pokémon's strength.

Research Question

What factors are associated with/help predict Pokemon strength?

Description Of The Data

The raw data set was obtained from kaggle.com. The data set can be found by searching for "Pokemon for Data Mining and Machine Learning."(https://www.kaggle.com/datasets/alopez247/pokemon) Our raw data and clean data can be found on separate files. The raw data includes 721 Pokémon for the first six generations. The data set includes 21 variables for each of the 721 Pokémon. After we cleaned our data set we used the variables HP, Attack, Defense, Sp_Atk, Sp_Def, Height_m, Weight_kg, Catch_Rate as our predictor variables and Total as our response variable. The scatter plot matrix below shows the significance of each variable and indication that they should be included in our models.



Methods

We cleaned the data by looking at scatter plots, bar charts, statistical summaries, and histograms. We narrowed down the predictor variables based on what we assumed are correlated with the Pokemon strength (Total). We had to collapse our categorical variables based on what we assumed were stronger types of Pokemon. Methods that were used to come up with the model are adjusted R^2, Cp, step wise, regressed subset, hypothesis test of the interaction terms.

Adjusted R^2 Method

```
##      X1    X2    X3    X4    X5    X6    X7    X8 num_in_model     adjr2
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE             1 0.5542006
## 2 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE             2 0.8080117
## 3 FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE             3 0.8835484
## 4  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE             4 0.9076524
## 5  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE             5 0.9314204
## 6  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE             6 0.9362168
## 7  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE             7 0.9394738
## 8  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE             8 0.9398687
```

Cp Method

```
##      X1    X2    X3    X4    X5    X6    X7    X8 num_in_model         Cp
## 1 FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE             1 4613.49536
## 2 FALSE  TRUE FALSE FALSE  TRUE FALSE FALSE FALSE             2 1577.44170
## 3 FALSE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE             3  675.55734
## 4  TRUE  TRUE FALSE  TRUE  TRUE FALSE FALSE FALSE             4  388.60811
## 5  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE FALSE             5  106.45574
## 6  TRUE  TRUE  TRUE  TRUE  TRUE FALSE FALSE  TRUE             6   50.36201
## 7  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE             7   12.68268
## 8  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE             8    9.00000
```

Stepwise Regression Method

```
##
##                                Stepwise Selection Summary
## ----------------------------------------------------------------------------------------
##                         Added/                   Adj.
## Step     Variable     Removed     R-Square    R-Square      C(p)          AIC         RMSE
## ----------------------------------------------------------------------------------------
##    1    Catch_Rate    addition       0.555       0.554    4613.4950     -373.9622    0.1862
##    2      Sp_Def      addition       0.690       0.689    3001.2360     -632.0506    0.1556
##    3      Attack      addition       0.846       0.845    1130.3140    -1135.5770    0.1096
##    4      Sp_Atk      addition       0.898       0.897     511.8890    -1429.4385    0.0894
##    5        HP        addition       0.917       0.916     289.7490    -1573.4183    0.0808
##    6     Defense      addition       0.937       0.936      50.3620    -1770.8914    0.0704
##    7    Weight_kg     addition       0.940       0.939      12.6830    -1807.6912    0.0686
##    8     Height_m     addition       0.941       0.940       9.0000    -1811.4228    0.0684
## ----------------------------------------
```

Model Results

We determined the best model by using the subset model selection method. Based on this model it has the highest Adjusted R^2 and the Cp score is the closest to the numbers of predictors of the model. Based on our best model below. Our predictor variables explain 96% of the Y variation. As Attack increases by 1, Total increases by 1.26 on average, etc.

Best Model

$Log(Total) = 49.73 + 0.82_{HP} + 1.26_{Attack} + 0.69_{Defense} + 1.35_{SpAtk} + 1.19_{SpDef} + 5.5_{Height\_m} - 0.014_{Weight\_kg} - 0.091_{Catch\_Rate} + 3.31_{Type1\_collapsedFire} - 1.88_{Type1\_collapsedGrass} + 6.05_{Type1\_collapsedOther} + 4.11_{Type1\_collapsedWater} - 0.0012_{SpAtk*SpDef} - 0.007_{Height*Weight}$

Discussion

According to our research the factors that best contribute to a Pokemon's strength are its Special Attack, Special Defense, Attack, Defense, Height, Weight and Catch Rate. Catch Rate is an inverse method, as the lower the catch rate the stronger the Pokemon. A Pokemon is also likely to be stronger if it is either Dragon or Grass type.

Limitations & Future Directions

One limitation that we came across were the multiple categories for our Type 1 and Type 2 variables. They each contain 18 unique elements. We overcame the issue by looking at which Pokémon had a higher total. We selected the five elements that are common in Type 1 and had higher totals. We have no future plans to work with the data set.