DATA SCIENCE
pokemon analysis

Team Member:林冠宇、黃子瑋、蔡政融

Feature Engineering

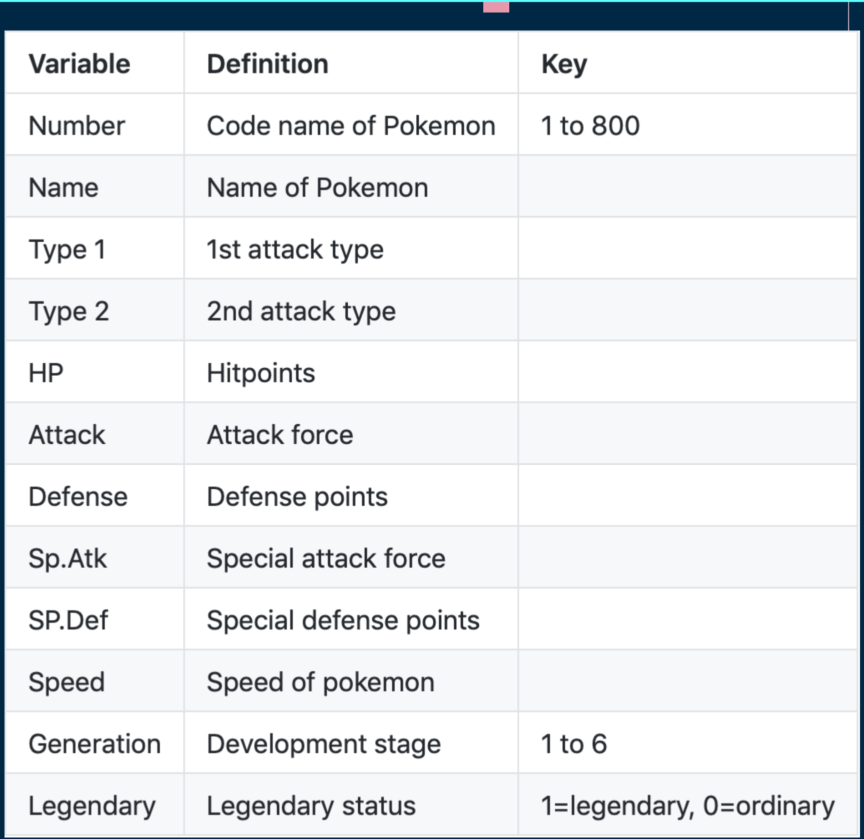
01

Input

→ Data source : kaggle dataset

→ Input format : csv file

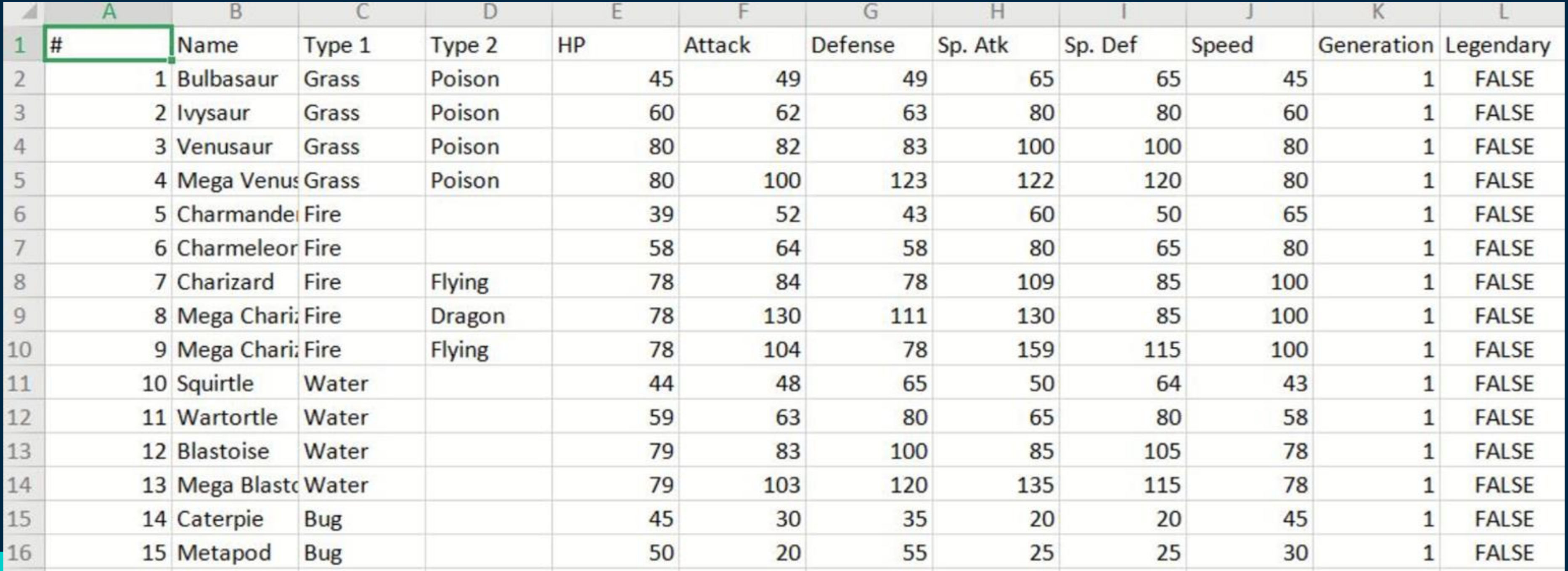
→ The data set contains 12 variables



Variable	Definition	Key
Number	Code name of Pokemon	1 to 800
Name	Name of Pokemon	
Type 1	1st attack type	
Type 2	2nd attack type	
HP	Hitpoints	
Attack	Attack force	
Defense	Defense points	
Sp.Atk	Special attack force	
SP.Def	Special defense points	
Speed	Speed of pokemon	
Generation	Development stage	1 to 6
Legendary	Legendary status	1=legendary, 0=ordinary

Feature Engineering

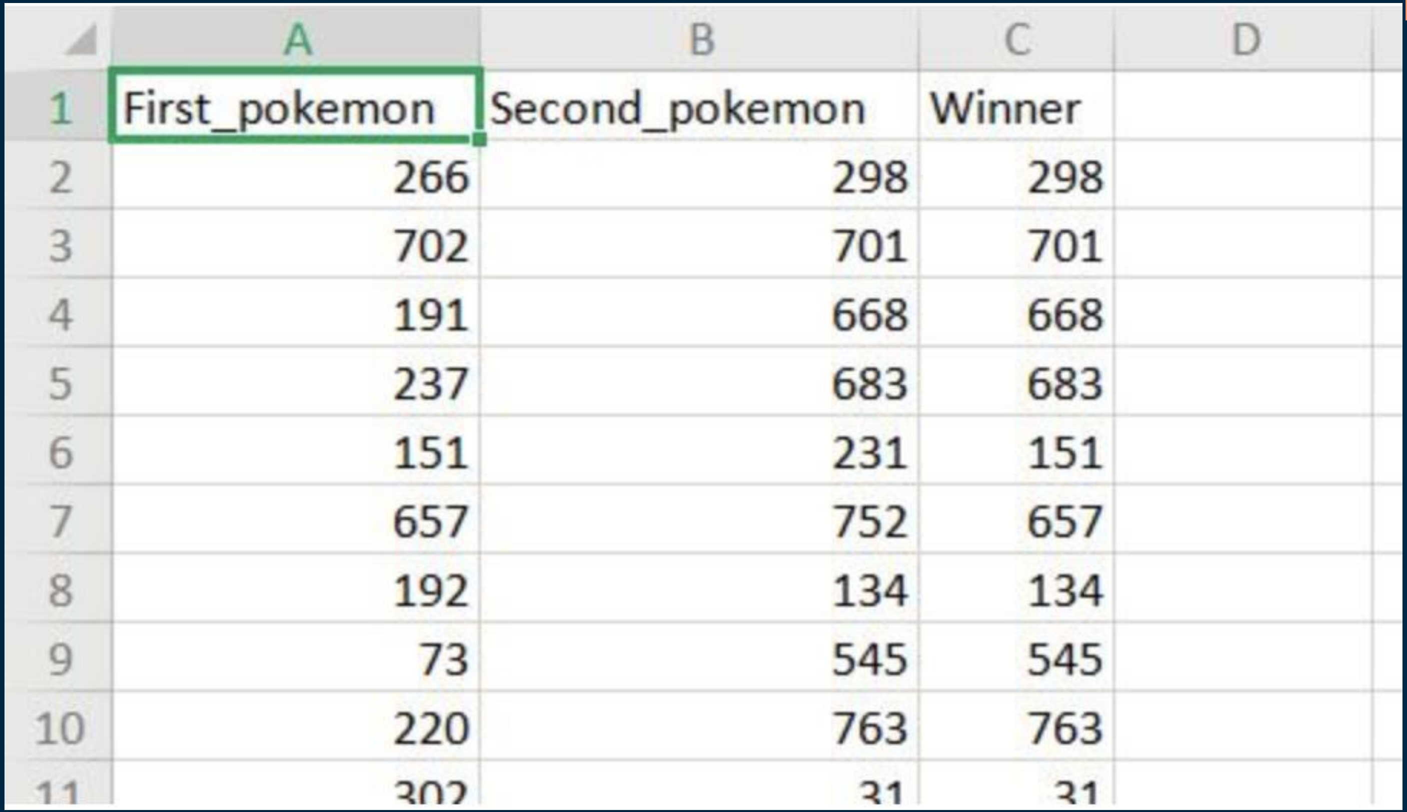
pokemon.csv



	A	B	C	D	E	F	G	H	I	J	K	L
1	#	Name	Type 1	Type 2	HP	Attack	Defense	Sp. Atk	Sp. Def	Speed	Generation	Legendary
2	1	Bulbasaur	Grass	Poison	45	49	49	65	65	45	1	FALSE
3	2	Ivysaur	Grass	Poison	60	62	63	80	80	60	1	FALSE
4	3	Venusaur	Grass	Poison	80	82	83	100	100	80	1	FALSE
5	4	Mega Venus	Grass	Poison	80	100	123	122	120	80	1	FALSE
6	5	Charmande	Fire		39	52	43	60	50	65	1	FALSE
7	6	Charmeleor	Fire		58	64	58	80	65	80	1	FALSE
8	7	Charizard	Fire	Flying	78	84	78	109	85	100	1	FALSE
9	8	Mega Chari	Fire	Dragon	78	130	111	130	85	100	1	FALSE
10	9	Mega Chari	Fire	Flying	78	104	78	159	115	100	1	FALSE
11	10	Squirtle	Water		44	48	65	50	64	43	1	FALSE
12	11	Wartortle	Water		59	63	80	65	80	58	1	FALSE
13	12	Blastoise	Water		79	83	100	85	105	78	1	FALSE
14	13	Mega Blastc	Water		79	103	120	135	115	78	1	FALSE
15	14	Caterpie	Bug		45	30	35	20	20	45	1	FALSE
16	15	Metapod	Bug		50	20	55	25	25	30	1	FALSE

Feature Engineering

combats.csv



First_pokemon	Second_pokemon	Winner
266	298	298
702	701	701
191	668	668
237	683	683
151	231	151
657	752	657
192	134	134
73	545	545
220	763	763
302	31	31

Feature Processing

- target
  - Calculate "Win rate" : We record the number of games and divide the individual wins by the total number of games played to calculate the win rate.
  - "Mega" variable creation : We have specially selected whether it is Mega Pokemon for EDA to see if it will have an impact on victory.

Feature engineering

→ Missing value

- In our dataset, three pokemon's name are missing.
- There is a Pokemon that is divided into male and female and has a gender symbol next to the name label, which causes an error when reading the dataset. We also fixed this problem manually.

Model Selection

02

Model Selection

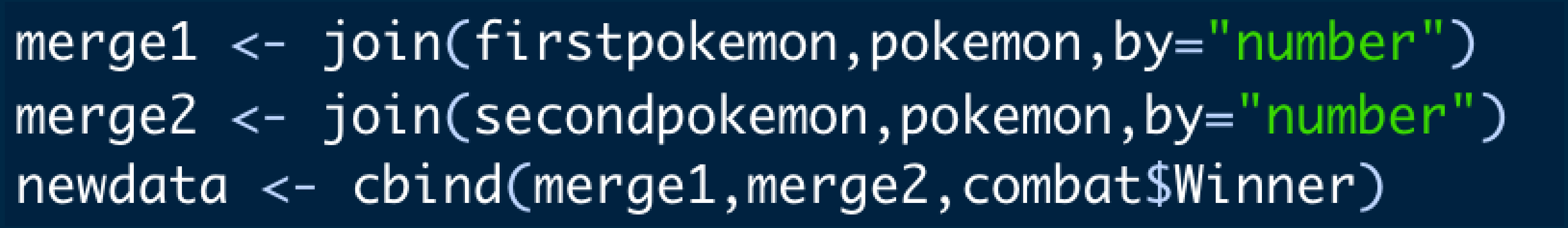
→ Data transition

→ 3 models

→ Evaluation

Data transition

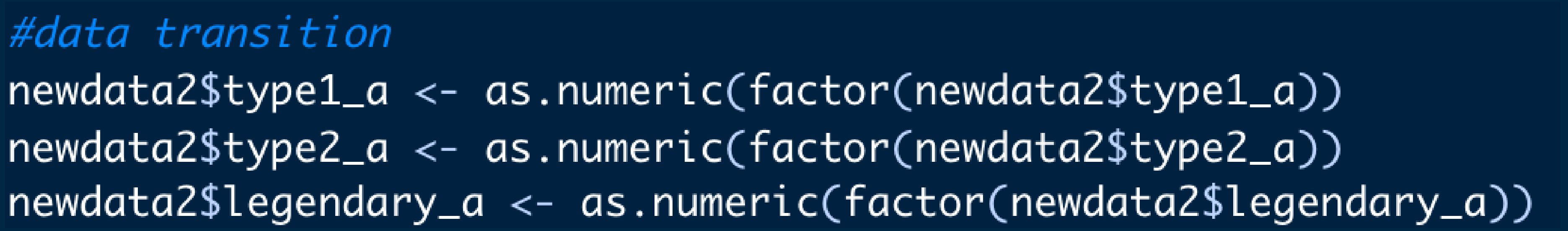
→ Merge two input files



```r
merge1 <- join(firstpokemon,pokemon,by="number")
merge2 <- join(secondpokemon,pokemon,by="number")
newdata <- cbind(merge1,merge2,combat$Winner)
```

Data transition

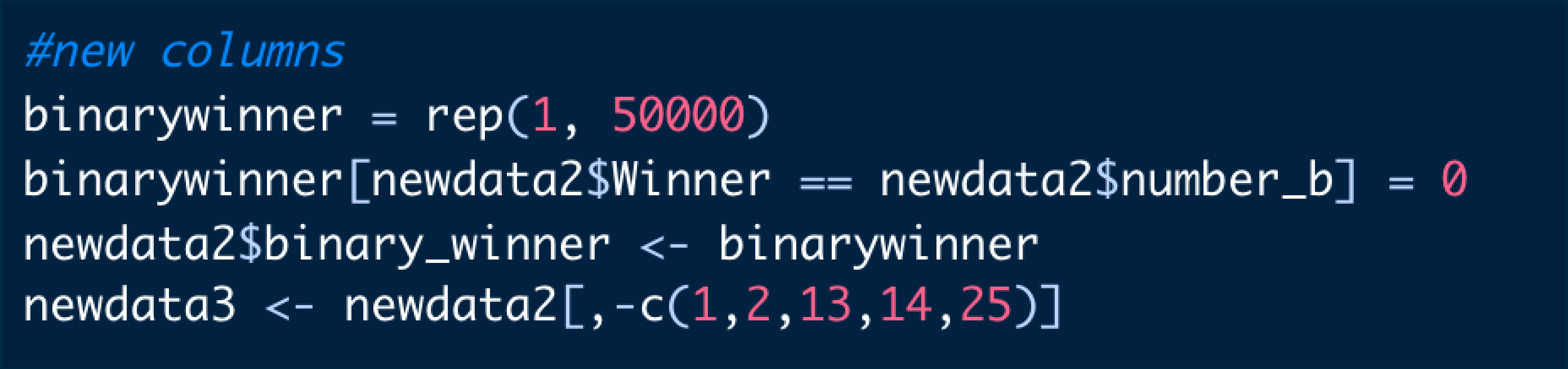
- ➔ From character to numeric



```
#data transition
newdata2$type1_a <- as.numeric(factor(newdata2$type1_a))
newdata2$type2_a <- as.numeric(factor(newdata2$type2_a))
newdata2$legendary_a <- as.numeric(factor(newdata2$legendary_a))
```

Data transition

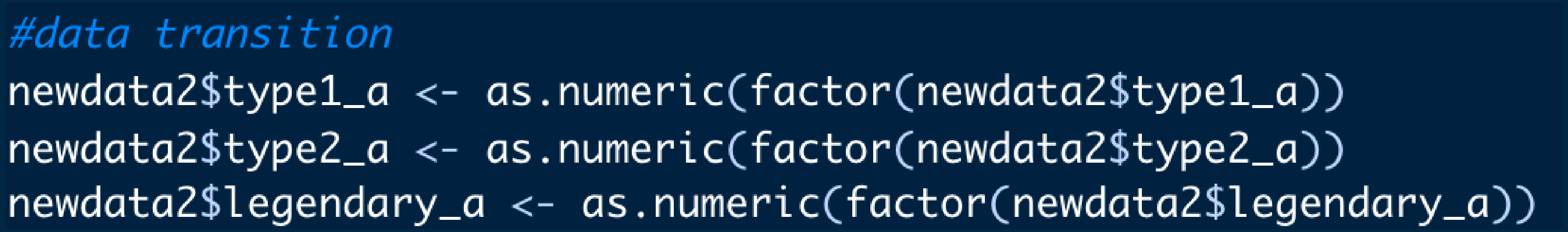
→ Delete unnecessary columns and create a column called “binarywinner” as label



```
#new columns
binarywinner = rep(1, 50000)
binarywinner[newdata2$Winner == newdata2$number_b] = 0
newdata2$binary_winner <- binarywinner
newdata3 <- newdata2[,-c(1,2,13,14,25)]
```

Data transition

→ From character to numeric



```
#data transition
newdata2$type1_a <- as.numeric(factor(newdata2$type1_a))
newdata2$type2_a <- as.numeric(factor(newdata2$type2_a))
newdata2$legendary_a <- as.numeric(factor(newdata2$legendary_a))
```

Models

→ Logistic regression

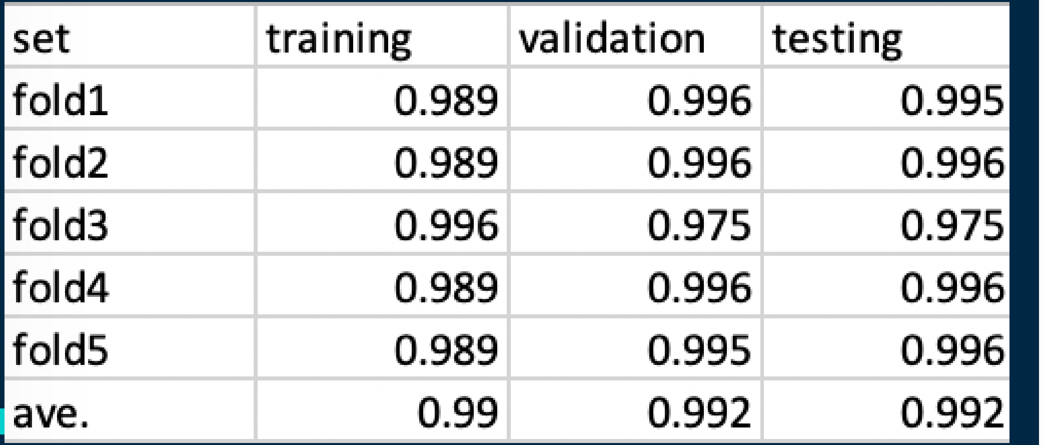
→ Random Forest

→ XGboost

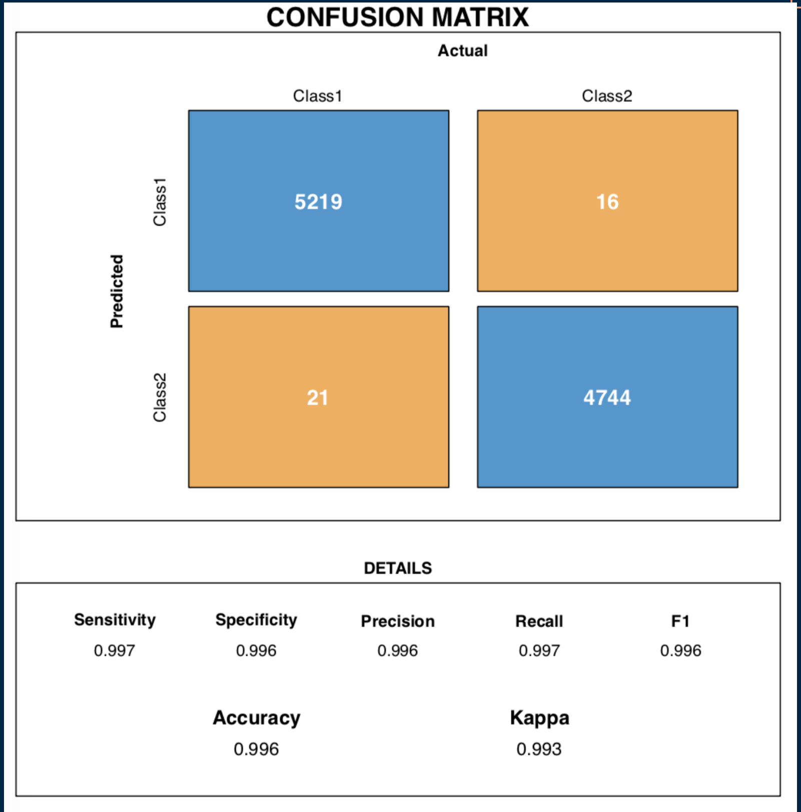
Evaluation

→ Random Forest as example

- ◆ Csv & confusion matrix



set	training	validation	testing
fold1	0.989	0.996	0.995
fold2	0.989	0.996	0.996
fold3	0.996	0.975	0.975
fold4	0.989	0.996	0.996
fold5	0.989	0.995	0.996
ave.	0.99	0.992	0.992



CONFUSION MATRIX

		Actual	
		Class1	Class2
Predicted	Class1	5219	16
	Class2	21	4744

DETAILS

Sensitivity	Specificity	Precision	Recall	F1
0.997	0.996	0.996	0.997	0.996

Accuracy	Kappa
0.996	0.993

Demo

03

https://ziweihuang.shinyapps.io/FinalProject/

Improvements 04

Bonus

Mega Rayquaza

Total battle numbers: 134 games(10 with same opponents)

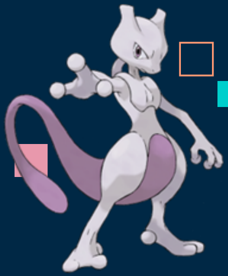
Win rate: 94.78% (win 127 games, loss 7 games)

Question:

According to the conclusion of our EDA, "Mega Rayquaza" is a Pokémon that ranks in the top ten in total ability, but why the win rate does not even ranks in the top ten?

Improvements

Lose to Mewtwo 超夢
(Normal & Mega-X & Mega-Y)



Lose to Cleffa 皮皮寶 twice

Lose to Florges 花傑夫人

Lose to Floette 花葉蒂

Improvements

- Among them, Mewtwo is "Legendary" Pokémon, and the total ability is similar to "Mega Rayquaza ".
- The other three Pokémon are general Pokémon (non-legendary), and their win rate is not high.
- Win rate:
  - Cleffa 皮寶寶: 6.25%
  - Florges 花傑夫人: 62.81%
  - Floette 花葉蒂: 39.67%

  They are all "Fairy" Pokémon !

  They are only the 3 Fairy Pokémon that Mega Rayquaza Battle

  ** Note: Florges花傑夫人 is the evolutionary pattern of Floette花葉蒂

Improvements

- Facing the legendary Pokémon (this refers to Mewtwo):
  - A little lower total ability may be a reason to loss
  - Mewtwo is a "Psychic" Pokemon, the Mega Rayquaza is a "Dragon + Flying" Pokemon, and there is no advantage or disadvantage in the type.
- Facing other general Pokémon:
  - "Fairy" restrains "Dragon" (Fairy moves attack Dragon and damage multiplied by 2)
  - In addition, after detailed exploration of skill moves, we found that all three Pokémon have moves that reduce the ability of opponents.
  - E.g : Cleffa 's "Angel Kiss" 天使之吻 effect is "Make opponents confused"
  - The "Mist Field" 薄霧場地 of Florges and Floette effect is "Make Dragon-type moves reduced by 50% damage"

Improvements

- Conclusion :
    - In addition to total ability, type restraint is also an important factor to consider.
    - Besides, there are also moves to be consider. If we can collect data or obtain relevant data from other data sets, we think that from the previous example, the prediction results may be improved.

THANK you for listening