Midterm Project Report

Yihan Feng (yf2555)

Introduction

Pokemon are fictional creatures, which can be captured by players and trained to battle each other for sport games. The dataset I used for this project includes basic information of 721 unique Pokemon, such as their names, battle features (health point, attack value, defense value, attack speed, etc), types, egg group, color, height, and weight, etc. Legendary Pokemon are a group of incredibly rare and often powerful Pokemon, which is also included in this dataset. As every player dreams to have a legendary Pokemon, it also leads to the question I am interested in: **Can we use the basic information to predict whether the Pokemon would be legendary?**
(Dataset Source: https://www.kaggle.com/alopez247/pokemon)

The dataset overall is tidy. I removed the variable “number”, which just lists the numbers of Pokemon. By observation, I found that the variable “total” looks like the sum of “hp”, “attack”, “defense”, “speed”, “sp_atk”, and “sp_def”, and I calculated with all values in the columns and proved my assumption. However, we still need further visualization or analysis to determine whether we should keep the “total” or the six battle features specifically. During visualization coding, I found that there is one line with all it’s values in wrong columns, and I removed it from the dataset.

Exploratory Analysis

I used contingency table (Figure 1) to show the overall relationship between legendary and non-legendary Pokemon. According to the table, I observed that legendary Pokemon have much higher mean and median than non-legendary ones in total battle features, health point, attack, defense, speed, attack speed, defense speed, proportion of not having gender, undiscovered egg groups, no mega evolution, height, weight, and low catch rate. While other variables do not have significant differences on mean, median, or proportions.

I also used two histograms to show the distribution of legendary Pokemon in total battle features (Figure 2) and 6 specific battle features (Figure 3). From Figure 1, I observed that all of the legendary Pokemon have over 550 total battle features. From Figure 2, I observed that most legendary Pokemon have each of the battle features values greater than average, but they are not showing an absolute trend or line. Therefore, I decided to use total battle features for future modeling.

Model

Based on the exploratory analysis, I decided to use the continuous and binary predictors that are significantly different between legendary and non-legendary Pokemon, which are:
* `is_legendary`: the response/outcome, whether the Pokemon is legendary or not (Yes/No).
* `total`: The sum of 6 battle features values.
* `has_gender`: whether the Pokemon has a gender (True/False).
* `has_mega_evolution`: whether the Pokemon has mega evolution (True/False).
* `height_m`: the height of Pokemon in Meters.
* `weight_kg`: the weight of Pokemon in Kilograms.
* `catch rate`: the officially set catch rate of each Pokemon (range from 0 ~ 255).

As the response is a binary categorical variable, I used classification to build models, which are LDA, QDA, and Naive Bayes methods, and GAM from logistic regression. I assume all of the subjects are independent in each class. And for LDA method, I assume the subjects are normally distributed. I divided the dataset 70% to training set, and the left 30% to testing set.

`resamples` function is used to compare the four training models, and to select a better one from summary table (Figure 4). In the ROC section, four models have high values, ranged from 0.99 ~ 1. Among the four models, GAM model gives the highest value, and we consider it performs better.

`predict` and `roc` functions are used to test, evaluate, and compare the four testing sets, and to select a better one from the ROC and AUC plot (Figure 5). According to the graph and AUC values, four sets have high values, ranged from 0.97 ~ 0.99. Among the four sets, Naive Bayes set gives the highest value, and GAM set follows.

Conclusion

Based on the performance of training and testing, I would select GAM as the final model for future prediction. Although Naive Bayes set performs the best in testing, the result contradicts with the NB assumption; it is useful when p is very large, while the p of legendary is only about 0.06. So I would not consider it as the final model. *GAM model performs good in both training and testing, and I would consider it as the final model.* This result corresponds to my expectation that the final model would be GAM or QDA. LDA is better for K > 2, while K = 2 in our dataset. And the reason not expecting Naive Bayes method is explained before.

Because of my limited knowledge about machine learning, I removed some of the predictors (multi-level categorical variables) from the training and testing sets, which is one of the major limitation of this project. Better Approaches might exist if we have include all meaningful predictors.

Appendix

All Codes are in the “yf2555_mtp.Rmd” file.

Figure 1. Contingency Table of Legendary and Non-legendary Pokemon

	False (N=674)	True (N=46)	Total (N=720)
type_1			
Bug	63 (9.3%)	0 (0.0%)	63 (8.8%)
Dark	26 (3.9%)	2 (4.3%)	28 (3.9%)
Dragon	17 (2.5%)	7 (15.2%)	24 (3.3%)
Electric	33 (4.9%)	3 (6.5%)	36 (5.0%)
Fairy	16 (2.4%)	1 (2.2%)	17 (2.4%)
Fighting	25 (3.7%)	0 (0.0%)	25 (3.5%)
Fire	42 (6.2%)	5 (10.9%)	47 (6.5%)
Flying	2 (0.3%)	1 (2.2%)	3 (0.4%)
Ghost	22 (3.3%)	1 (2.2%)	23 (3.2%)
Grass	64 (9.5%)	2 (4.3%)	66 (9.2%)
Ground	28 (4.2%)	2 (4.3%)	30 (4.2%)
Ice	21 (3.1%)	2 (4.3%)	23 (3.2%)
Normal	91 (13.5%)	2 (4.3%)	93 (12.9%)
Poison	27 (4.0%)	0 (0.0%)	27 (3.8%)
Psychic	39 (5.8%)	8 (17.4%)	47 (6.5%)
Rock	38 (5.6%)	3 (6.5%)	41 (5.7%)
Steel	18 (2.7%)	4 (8.7%)	22 (3.1%)
Water	102 (15.1%)	3 (6.5%)	105 (14.6%)
type_2			
	351 (52.1%)	19 (41.3%)	370 (51.4%)
Bug	3 (0.4%)	0 (0.0%)	3 (0.4%)
Dark	16 (2.4%)	0 (0.0%)	16 (2.2%)
Dragon	11 (1.6%)	3 (6.5%)	14 (1.9%)
Electric	5 (0.7%)	1 (2.2%)	6 (0.8%)
Fairy	17 (2.5%)	1 (2.2%)	18 (2.5%)
Fighting	16 (2.4%)	3 (6.5%)	19 (2.6%)
Fire	7 (1.0%)	2 (4.3%)	9 (1.2%)
Flying	78 (11.6%)	9 (19.6%)	87 (12.1%)
Ghost	11 (1.6%)	1 (2.2%)	12 (1.7%)
Grass	18 (2.7%)	0 (0.0%)	18 (2.5%)
Ground	29 (4.3%)	1 (2.2%)	30 (4.2%)
Ice	9 (1.3%)	1 (2.2%)	10 (1.4%)
Normal	4 (0.6%)	0 (0.0%)	4 (0.6%)
Poison	31 (4.6%)	0 (0.0%)	31 (4.3%)
Psychic	24 (3.6%)	3 (6.5%)	27 (3.8%)
Rock	14 (2.1%)	0 (0.0%)	14 (1.9%)
Steel	18 (2.7%)	1 (2.2%)	19 (2.6%)
Water	12 (1.8%)	1 (2.2%)	13 (1.8%)
total			
Mean (SD)	404.4 (98.6)	620.2 (45.0)	418.1 (109.6)
Median (Q1, Q3)	410.0 (316.0, 490.0)	600.0 (580.0, 677.5)	424.0 (320.0, 499.2)
hp			
Mean (SD)	66.7 (25.3)	94.0 (18.8)	68.4 (25.9)
Median (Q1, Q3)	65.0 (50.0, 78.0)	91.0 (80.0, 103.8)	65.0 (50.0, 80.0)
attack			
Mean (SD)	72.7 (27.9)	108.6 (23.5)	75.0 (29.0)

	False (N=674)	True (N=46)	Total (N=720)
Median (Q1, Q3)	70.0 (52.0, 90.0)	107.5 (90.0, 123.8)	74.5 (53.0, 95.0)
defense			
Mean (SD)	68.8 (28.3)	101.6 (26.5)	70.9 (29.3)
Median (Q1, Q3)	65.0 (50.0, 84.0)	100.0 (90.0, 118.8)	65.0 (50.0, 85.0)
sp_atk			
Mean (SD)	65.7 (26.3)	113.9 (25.4)	68.8 (28.8)
Median (Q1, Q3)	61.0 (45.0, 85.0)	117.5 (96.2, 130.0)	65.0 (45.0, 90.0)
sp_def			
Mean (SD)	66.8 (25.1)	106.5 (27.2)	69.3 (27.0)
Median (Q1, Q3)	65.0 (50.0, 80.0)	100.0 (90.0, 120.0)	65.0 (50.0, 85.0)
speed			
Mean (SD)	63.7 (26.5)	95.6 (20.4)	65.7 (27.3)
Median (Q1, Q3)	60.0 (45.0, 82.8)	99.0 (90.0, 108.0)	65.0 (45.0, 85.0)
generation			
Mean (SD)	3.3 (1.7)	3.8 (1.5)	3.3 (1.7)
Median (Q1, Q3)	3.0 (2.0, 5.0)	4.0 (3.0, 5.0)	3.0 (2.0, 5.0)
color			
Black	29 (4.3%)	3 (6.5%)	32 (4.4%)
Blue	125 (18.5%)	9 (19.6%)	134 (18.6%)
Brown	105 (15.6%)	5 (10.9%)	110 (15.3%)
Green	74 (11.0%)	5 (10.9%)	79 (11.0%)
Grey	65 (9.6%)	4 (8.7%)	69 (9.6%)
Pink	39 (5.8%)	2 (4.3%)	41 (5.7%)
Purple	61 (9.1%)	3 (6.5%)	64 (8.9%)
Red	70 (10.4%)	5 (10.9%)	75 (10.4%)
White	48 (7.1%)	4 (8.7%)	52 (7.2%)
Yellow	58 (8.6%)	6 (13.0%)	64 (8.9%)
has_gender			
False	37 (5.5%)	40 (87.0%)	77 (10.7%)
True	637 (94.5%)	6 (13.0%)	643 (89.3%)
pr_male			
Mean (SD)	0.6 (0.2)	0.8 (0.4)	0.6 (0.2)
Median (Q1, Q3)	0.5 (0.5, 0.5)	1.0 (0.6, 1.0)	0.5 (0.5, 0.5)
egg_group_1			
Amorphous	41 (6.1%)	0 (0.0%)	41 (5.7%)
Bug	66 (9.8%)	0 (0.0%)	66 (9.2%)
Ditto	1 (0.1%)	0 (0.0%)	1 (0.1%)
Dragon	10 (1.5%)	0 (0.0%)	10 (1.4%)
Fairy	30 (4.5%)	0 (0.0%)	30 (4.2%)
Field	169 (25.1%)	0 (0.0%)	169 (23.5%)
Flying	44 (6.5%)	0 (0.0%)	44 (6.1%)
Grass	27 (4.0%)	0 (0.0%)	27 (3.8%)
Human-Like	37 (5.5%)	0 (0.0%)	37 (5.1%)
Mineral	46 (6.8%)	0 (0.0%)	46 (6.4%)
Monster	73 (10.8%)	0 (0.0%)	73 (10.1%)
Undiscovered	27 (4.0%)	46 (100.0%)	73 (10.1%)
Water_1	74 (11.0%)	0 (0.0%)	74 (10.3%)
Water_2	15 (2.2%)	0 (0.0%)	15 (2.1%)
Water_3	14 (2.1%)	0 (0.0%)	14 (1.9%)
egg_group_2			
	484 (71.8%)	46 (100.0%)	530 (73.6%)
Amorphous	8 (1.2%)	0 (0.0%)	8 (1.1%)

	False (N=674)	True (N=46)	Total (N=720)
Bug	2 (0.3%)	0 (0.0%)	2 (0.3%)
Dragon	35 (5.2%)	0 (0.0%)	35 (4.9%)
Fairy	17 (2.5%)	0 (0.0%)	17 (2.4%)
Field	30 (4.5%)	0 (0.0%)	30 (4.2%)
Flying	6 (0.9%)	0 (0.0%)	6 (0.8%)
Grass	32 (4.7%)	0 (0.0%)	32 (4.4%)
Human-Like	15 (2.2%)	0 (0.0%)	15 (2.1%)
Mineral	8 (1.2%)	0 (0.0%)	8 (1.1%)
Monster	1 (0.1%)	0 (0.0%)	1 (0.1%)
Water_1	13 (1.9%)	0 (0.0%)	13 (1.8%)
Water_2	8 (1.2%)	0 (0.0%)	8 (1.1%)
Water_3	15 (2.2%)	0 (0.0%)	15 (2.1%)
has_mega_evolution			
False	633 (93.9%)	41 (89.1%)	674 (93.6%)
True	41 (6.1%)	5 (10.9%)	46 (6.4%)
height_m			
Mean (SD)	1.1 (0.9)	2.4 (1.7)	1.1 (1.0)
Median (Q1, Q3)	0.9 (0.6, 1.4)	2.0 (1.5, 3.2)	1.0 (0.6, 1.4)
weight_kg			
Mean (SD)	46.9 (66.0)	201.8 (197.2)	56.8 (89.1)
Median (Q1, Q3)	25.7 (9.0, 55.5)	196.5 (56.5, 293.8)	28.0 (9.5, 61.0)
catch_rate			
Mean (SD)	106.4 (74.8)	6.7 (12.0)	100.1 (76.5)
Median (Q1, Q3)	75.0 (45.0, 190.0)	3.0 (3.0, 3.0)	65.0 (45.0, 180.0)
body_style			
bipedal_tailed	150 (22.3%)	8 (17.4%)	158 (21.9%)
bipedal_tailless	101 (15.0%)	8 (17.4%)	109 (15.1%)
four_wings	18 (2.7%)	0 (0.0%)	18 (2.5%)
head_arms	35 (5.2%)	4 (8.7%)	39 (5.4%)
head_base	30 (4.5%)	0 (0.0%)	30 (4.2%)
head_legs	17 (2.5%)	0 (0.0%)	17 (2.4%)
head_only	33 (4.9%)	1 (2.2%)	34 (4.7%)
insectoid	30 (4.5%)	0 (0.0%)	30 (4.2%)
multiple_bodies	15 (2.2%)	0 (0.0%)	15 (2.1%)
quadruped	122 (18.1%)	12 (26.1%)	134 (18.6%)
serpentine_body	26 (3.9%)	3 (6.5%)	29 (4.0%)
several_limbs	13 (1.9%)	0 (0.0%)	13 (1.8%)
two_wings	54 (8.0%)	9 (19.6%)	63 (8.8%)
with_fins	30 (4.5%)	1 (2.2%)	31 (4.3%)

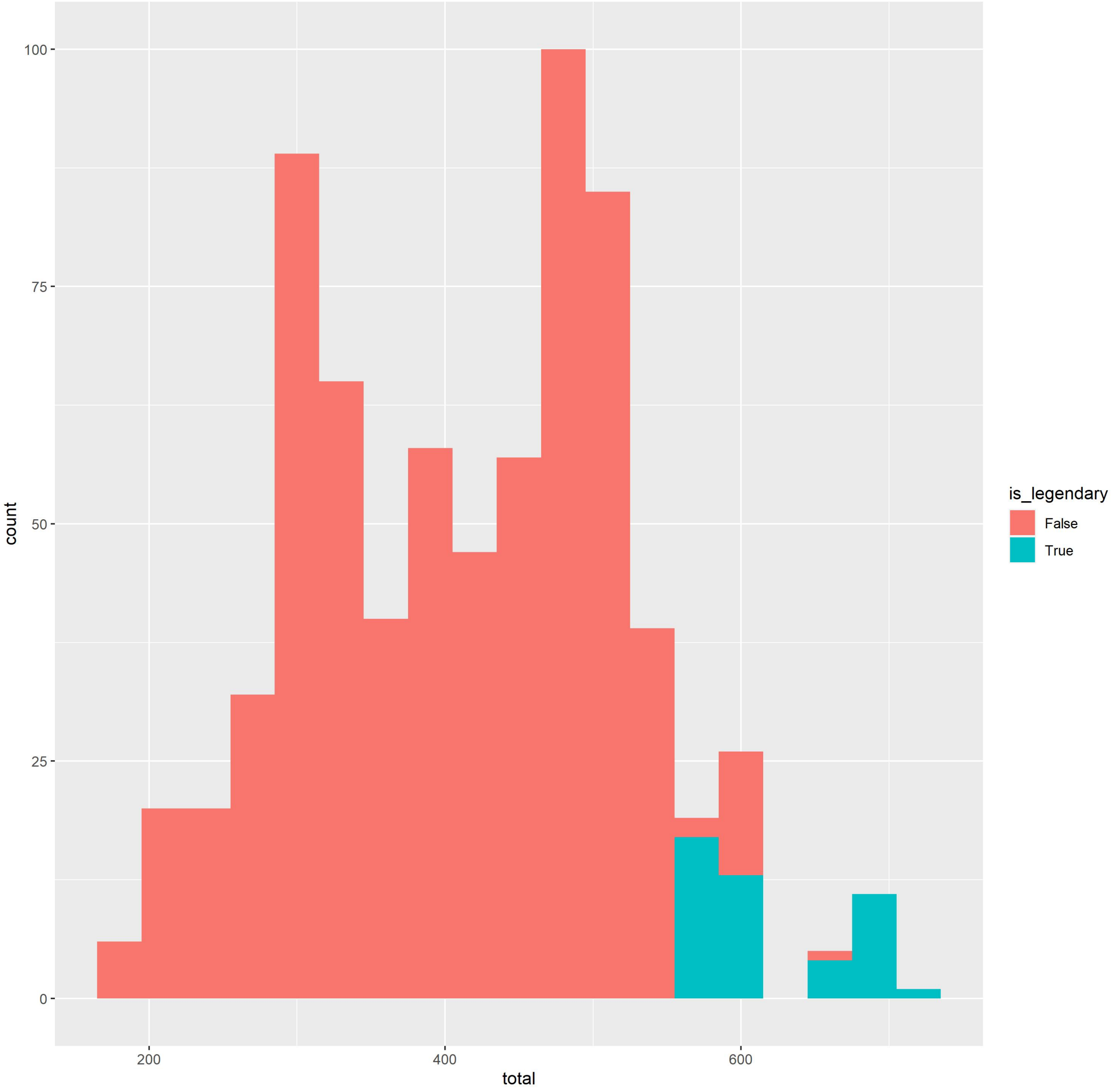


Figure 2. Histogram of total battle feature among legendary or non-legendary Pokemon

Figure 3. Histogram of 6 Features among legendary or non-legendary Pokemon

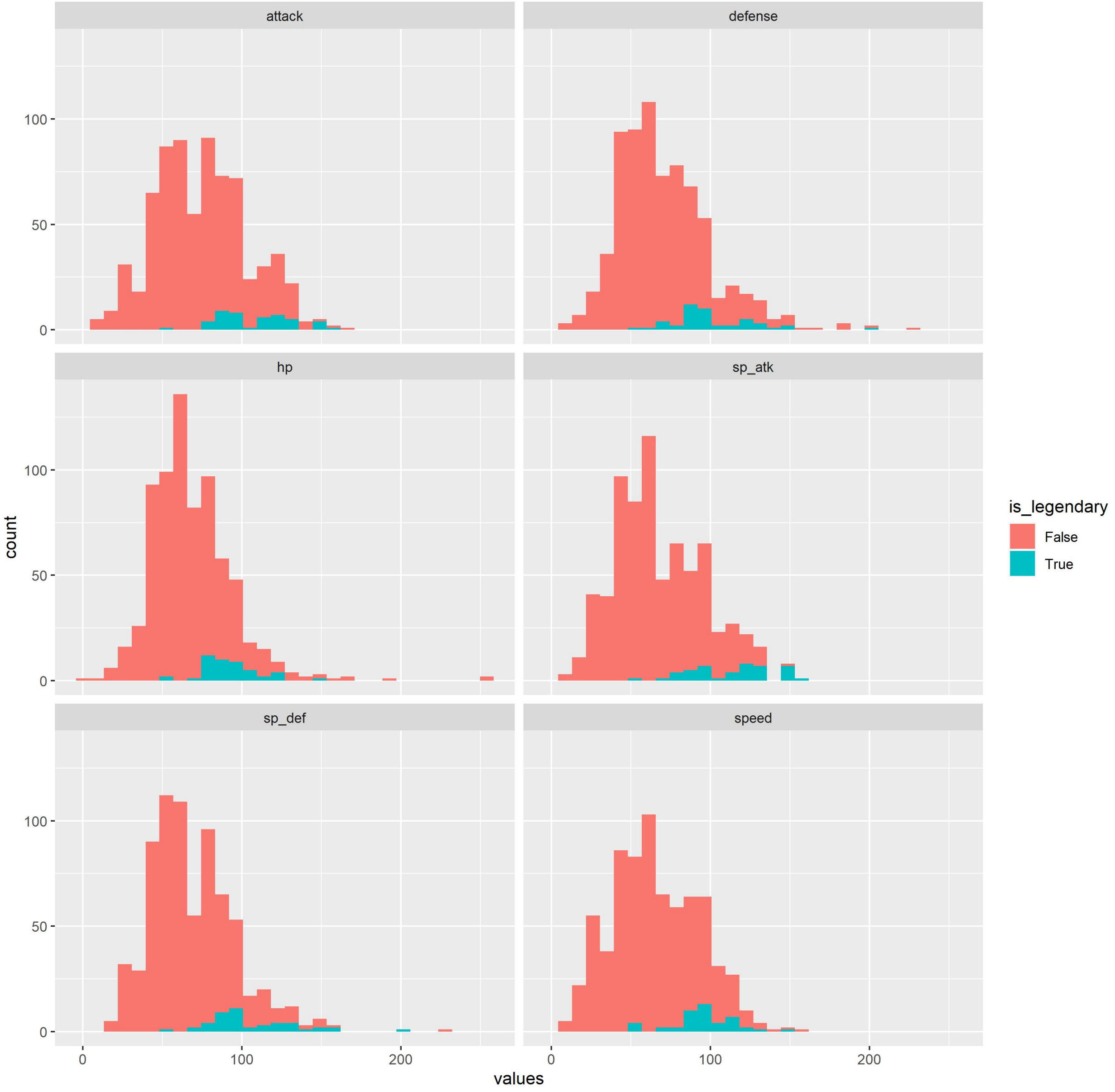
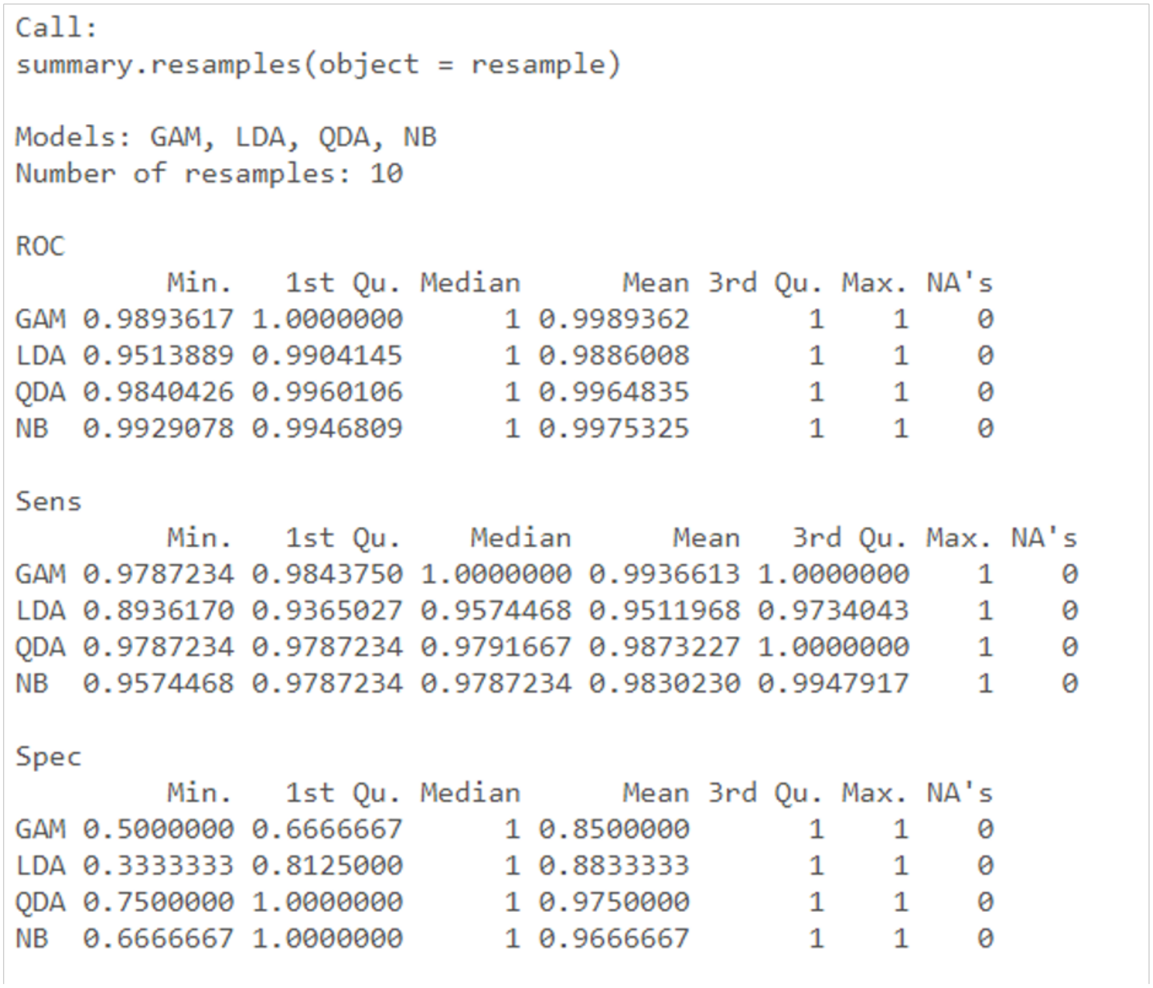
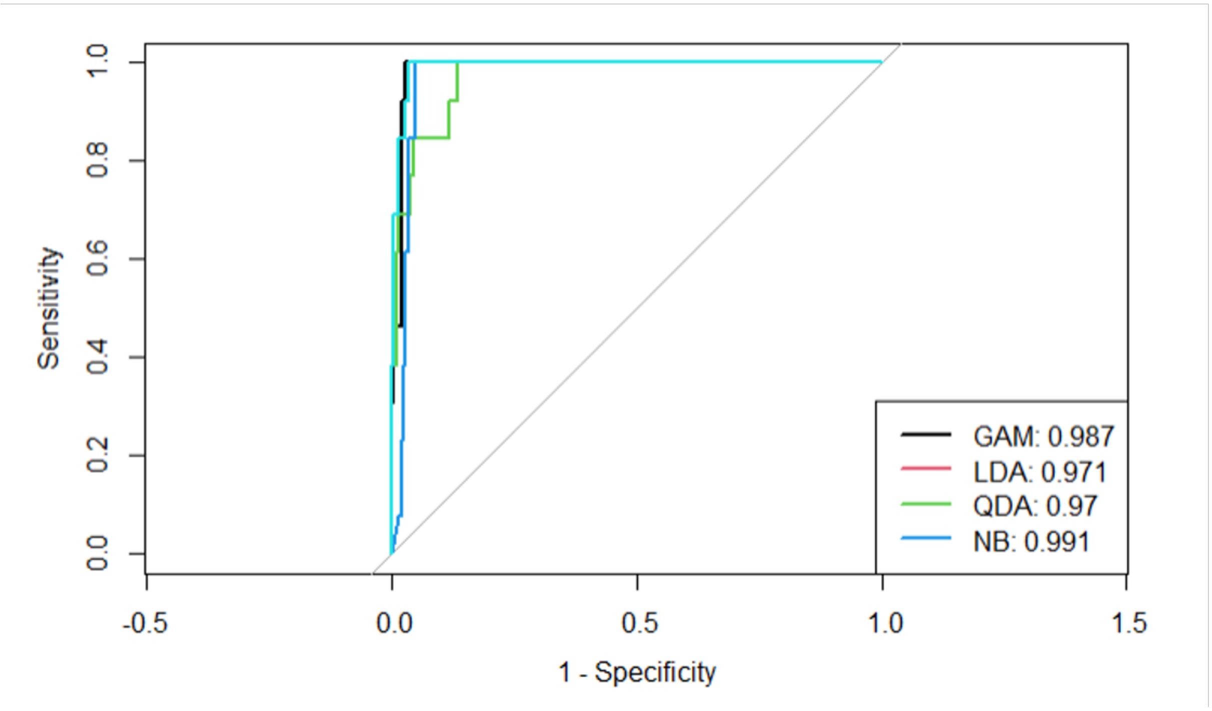


Figure 4. Resamples Summary Table of Training Models



```
Call:
summary.resamples(object = resample)

Models: GAM, LDA, QDA, NB
Number of resamples: 10

ROC
         Min.   1st Qu. Median      Mean 3rd Qu. Max. NA's
GAM 0.9893617 1.0000000      1 0.9989362       1    1    0
LDA 0.9513889 0.9904145      1 0.9886008       1    1    0
QDA 0.9840426 0.9960106      1 0.9964835       1    1    0
NB  0.9929078 0.9946809      1 0.9975325       1    1    0

Sens
         Min.   1st Qu.    Median      Mean   3rd Qu. Max. NA's
GAM 0.9787234 0.9843750 1.0000000 0.9936613 1.0000000    1    0
LDA 0.8936170 0.9365027 0.9574468 0.9511968 0.9734043    1    0
QDA 0.9787234 0.9787234 0.9791667 0.9873227 1.0000000    1    0
NB  0.9574468 0.9787234 0.9787234 0.9830230 0.9947917    1    0

Spec
         Min.   1st Qu. Median      Mean 3rd Qu. Max. NA's
GAM 0.5000000 0.6666667      1 0.8500000       1    1    0
LDA 0.3333333 0.8125000      1 0.8833333       1    1    0
QDA 0.7500000 1.0000000      1 0.9750000       1    1    0
NB  0.6666667 1.0000000      1 0.9666667       1    1    0
```

Figure 5. AUC Plot of Testing Sets


Sensitivity
1 - Specificity
GAM: 0.987
LDA: 0.971
QDA: 0.97
NB: 0.991