2022년 2학기
고급통계학특론I

Final Project

REPORT

222STG10 김희숙

목차

Abstract

게임에서 몬스터의 종족 레벨링은 플레이어 입장에서 매우 중요한 요소이다. 특히 캐릭터가 성장함에 따라 더욱 강한 몬스터를 만나게 되는 RPG 게임의 경우 최종 몬스터의 강함은 종족값 같은 물리적인 능력치 뿐만 아니라 복합적인 면에서 다른 몬스터들과 차별화 되어야한다. 예를 들어, RPG게임 포켓몬스터에서 최종적으로 등장하는 전설의 포켓몬은 플레이어의 게임 구매 여부를 결정하는 요소로 작용한다. 본 프로젝트는 6세대까지 등장한 포켓몬의 종족 데이터를 살펴본 후 전설의 포켓몬 여부를 분류하였다. 포켓몬 중 전설의 포켓몬의 비율이 작으므로 오버샘플링을 통해 데이터 불균형을 해결하고 분석을 수행하였다. 본 프로젝트는 차원 축소 여부에 따른 분류 머신러닝 모형을 비교하여 최적의 모형과 차원 축소 여부를 찾고자 하였다. 최종적으로 전설의 포켓몬 분류에 차원 축소는 적절해 보이지 않았으나 SIR, SAVE, DR 순서대로 Naive Bayes, SVM, XGBoost의 성능이 높게 나왔다.

---

1. 서론

RPG 게임은 플레이어에 해당하는 캐릭터를 주인공으로 모험을 하며 레벨을 올려 강한 적을 해치우며 계속 성장하는 것을 목표로 하는 게임의 일종이다. 이러한 RPG 게임 시스템에서 적, 즉 몬스터의 레벨과 능력치는 플레이어가 적을 해치울 수 있는가에 대한 지표로 게임을 즐기는 요소 중 하나이다. 일반적으로 몬스터의 등급은 종합 능력치로 구분되나, 최종 몬스터의 경우 능력치 외 게임마다 가진 특성에 기반한 복합적인 정보로 정의된다.

본 프로젝트에서는 RPG 게임 포켓몬스터 내 등장하는 포켓몬에 대한 정보 데이터를 통해 포켓몬의 능력치 및 속성 정보가 주어졌을 때, 게임의 메인 이라고 볼 수 있는 전설의 포켓몬 여부를 분류하고자 한다. 또한 feature의 개수가 10개가 넘어가는 포켓몬 데이터에 대해 dimension reduction을 적용하는 것이 분류 모델링을 함에 있어 효과가 있는지 알아보고자 한다.

2. 데이터 설명

본 프로젝트에 이용한 데이터는 'pokemon_alopez247'으로 kaggle[1]에서 제공하고 있는 데이터이다. RPG게임인 포켓몬스터 시리즈의 6세대까지 등장한 721마리 포켓몬 각각에 대한 능력치 및 캐릭터 정보를 포함하고있는 데이터이다. 포켓몬 중 전설의 포켓몬은 일반적인 보통 포켓몬보다 강하며[2] 대체로 각 세대마다 3가지 개체의 전설 포켓몬이 등장한다.

---

[1] 데이터 출처: https://www.kaggle.com/datasets/alopez247/pokemon

[2] 모든 전설의 포켓몬이 일반 포켓몬 보다 강한 능력치를 갖고 있는 것은 아니다. 일반 포켓몬 중에도 전설의 포켓몬과 비등하거나 더 강한 포켓몬이 존재한다. 예를 들어, 미끄래곤, 망나뇽, 한카리아스는 일반 포켓몬이며 라티오스, 비크니티, 쉐이미는 전설(/환상)의 포켓몬으로 모두 종합 능력치가 600으로 동일하다.

a. 데이터 셋 및 변수

본 프로젝트에서 사용한 데이터는 721개의 observation, 23개의 variable로 12개의 continuous variable과 11개의 categorical variable로 이루어져있다. 각 변수에 대한 자세한 설명은 [표 1]과 같다.

변수명	Detail	변수명	detail
Number	포켓몬 도감에 표기된 포켓몬 아이디 번호	isLegendary	전설의 포켓몬 여부
Name	포켓몬 이름	Color	포켓몬의 색
Type_1	포켓몬의 기본 타입	hasGender	성별(암컷/수컷) 유무 여부
Type_2	포켓몬의 두번째 타입	Pr_male	포켓몬의 성별이 수컷일 확률
Total	능력치의 합계 (Health Points, Attack, Defense, Special Attack, Special Defense, and Speed의 합계)	EggGroup1	포켓몬의 종 타입
HP	체력	EggGroup2	포켓몬의 두번째 종 타입
Attack	물리 공격력	hasMegaEvolution	메가 진화 유무 여부
Defense	물리 방어력	Height_m	포켓몬의 키(단위: meters)
Sp_Atk	특수 공격력	Weight_kg	포켓몬의 무게(단위: kg)
Sp_Def	특수 방어력	Catch_Rate	포켓몬 포획률(*숫자가 클 수록 잡기 쉬운 포켓몬을 의미)
Speed	속도	Body_Style	포켓몬 몸의 형태
Generation	포켓몬이 등장한 세대		

[표 1. 변수 설명]

b. 통계량 및 결측치

각 변수들의 통계량 및 summary는 [그림 1]과 같다. 연속형 변수의 경우 'Weight_kg'와 'Catch_Rate' 변수를 제외한 대부분의 변수가 mean과 median의 차이가 크지 않아 이상치는 없다고 판단되었으며, 범주형 변수의 경우 데이터 불균형이 보이는 변수들이 존재하였다. 특히 classification의 target인 'isLegendary'의 클래스가 675와 46으로, 비율이 14:1 정도로 데이터 불균형이 심각하였다. 모델링을 하는 과정에서 데이터 불균형을 해결하기 위한 resampling 방법이 필요하다고 판단된다. 데이터 내 결측치는 'Pr_Male'변수에만 77개 존재 했는데, 이는 성별을 갖지 않는 포켓몬(hasGender = False)의 개수와 일치했으며, 이에 대한 NA값으로 판단된다. 따라서, 'Pr_Male'의 NA값에 확률값 범위 밖인 1.5로 값을 대체하고 프로젝트를 진행한다.

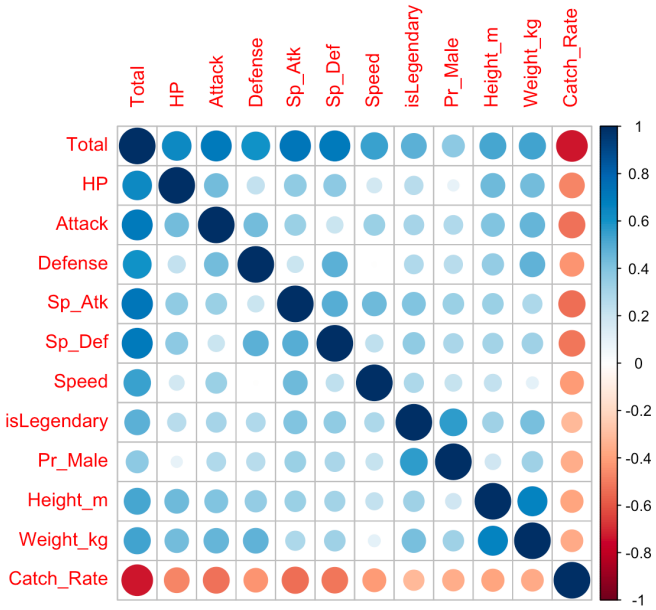
```
    Number            Name         Type_1         Type_2        Total      
Min.   :  1   Abomasnow :  1   Water  :105          :371   Min.   :180.0  
1st Qu.:181   Abra      :  1   Normal : 93   Flying  : 87   1st Qu.:320.0  
Median :361   Absol     :  1   Grass  : 66   Poison  : 31   Median :424.0  
Mean   :361   Accelgor  :  1   Bug    : 63   Ground  : 30   Mean   :417.9  
3rd Qu.:541   Aegislash :  1   Fire   : 47   Psychic : 27   3rd Qu.:499.0  
Max.   :721   Aerodactyl:  1   Psychic: 47   Fighting: 19   Max.   :720.0  
              (Other)   :715   (Other):300   (Other) :156                  
       HP             Attack          Defense          Sp_Atk           Sp_Def      
Min.   :  1.00   Min.   :  5.00   Min.   :  5.00   Min.   : 10.00   Min.   : 20.00  
1st Qu.: 50.00   1st Qu.: 53.00   1st Qu.: 50.00   1st Qu.: 45.00   1st Qu.: 50.00  
Median : 65.00   Median : 74.00   Median : 65.00   Median : 65.00   Median : 65.00  
Mean   : 68.38   Mean   : 75.01   Mean   : 70.81   Mean   : 68.74   Mean   : 69.29  
3rd Qu.: 80.00   3rd Qu.: 95.00   3rd Qu.: 85.00   3rd Qu.: 90.00   3rd Qu.: 85.00  
Max.   :255.00   Max.   :165.00   Max.   :230.00   Max.   :154.00   Max.   :230.00  

     Speed          Generation    isLegendary     Color     hasGender    
Min.   :  5.00   Min.   :1.000   False:675   Blue   :134   False: 77  
1st Qu.: 45.00   1st Qu.:2.000   True : 46   Brown  :110   True :644  
Median : 65.00   Median :3.000               Green  : 79              
Mean   : 65.71   Mean   :3.323               Red    : 75              
3rd Qu.: 85.00   3rd Qu.:5.000               Grey   : 69              
Max.   :160.00   Max.   :6.000               Purple : 65              
                                             (Other):189              
   Pr_Male            Egg_Group_1      Egg_Group_2  hasMegaEvolution
Min.   :0.0000   Field       :169            :530   False:675       
1st Qu.:0.5000   Monster     : 74   Dragon    : 35   True : 46       
Median :0.5000   Water_1     : 74   Grass     : 32                   
Mean   :0.5534   Undiscovered: 73   Field     : 31                   
3rd Qu.:0.5000   Bug         : 66   Fairy     : 17                   
Max.   :1.0000   Mineral     : 46   Human-Like: 15                   
NA's   :77       (Other)     :219   (Other)   : 61                   
   Height_m        Weight_kg        Catch_Rate              Body_Style   
Min.   : 0.100   Min.   :  0.10   Min.   :  3.0   bipedal_tailed  :158  
1st Qu.: 0.610   1st Qu.:  9.40   1st Qu.: 45.0   quadruped       :135  
Median : 0.990   Median : 28.00   Median : 65.0   bipedal_tailless:109  
Mean   : 1.145   Mean   : 56.77   Mean   :100.2   two_wings       : 63  
3rd Qu.: 1.400   3rd Qu.: 61.00   3rd Qu.:180.0   head_arms       : 39  
Max.   :14.500   Max.   :950.00   Max.   :255.0   head_only       : 34  
                                                 (Other)         :183  
```

[그림 1. 변수 통계량]

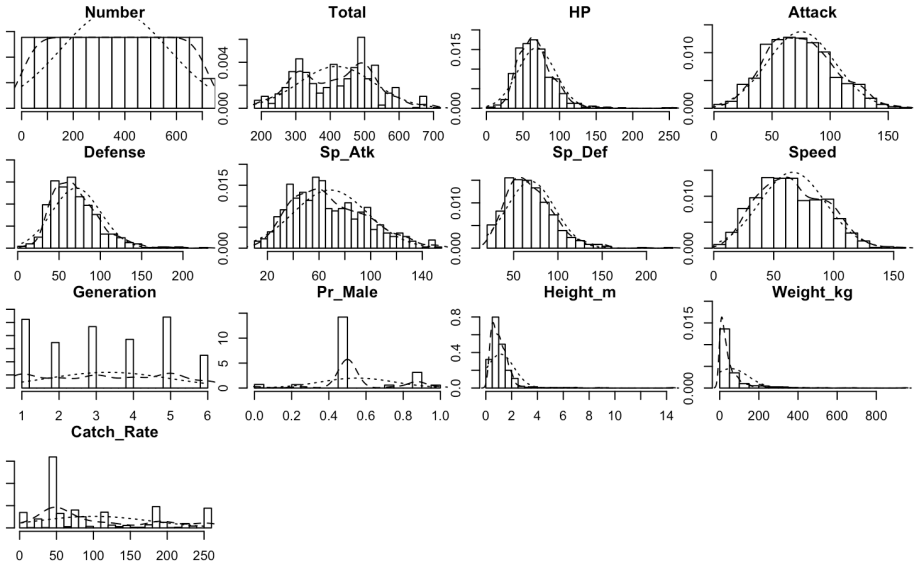
c. 상관관계 및 변수 분포

범주형 변수를 제외한 연속형 변수들의 correlation을 확인해보면 [그림 2]와 같다. correlation heatmap을 보면 몇 변수들 간 높은 correlation이 존재하는 것을 확인 할 수 있다. dimension reduction 방법 중 ols와 mave를 사용할 경우 regression 기반이라 변수 drop이 필수적으로 이루어 져야하나, 본 프로젝트에서 이용할 dimension reduction은 SIR, SAVE, DR 이므로 변수 drop은 진행하지 않아도 된다.

마찬가지로 연속형 변수들의 히스토그램을 확인해보면 [그림 3]과 같다. 데이터의 값이 정수이면서 범위가 좁은 변수들은 히스토그램의 간격이 넓게 나온 것을 확인할 수 있다. 대부분의 변수들이 정규 분포의 모양을 보이고 있으나, 'Height_m', 'Weight_kg', 'Catch_Rate'는 오른쪽으로 꼬리가 긴 형태를 갖는 것을 확인할 수 있다. 이상치를 확인해 본 결과 'Height_m', 'Weight_kg' 두 변수에서 이상치가 존재하였으며, dimension reduction 방법중 SAVE의 경우 이상치에 민감하므로 두 변수에 대해 공통으로 이상치인 데이터를 제거하였다. 이상치로 판단된 observation은 총 24개이며, 'isLegendary' 변수의 클래스 값은 'False: 667'과 'True: 30'으로 변화하였다.



[그림 2. correlation plot]



[그림 3. histogram plot]

3. 데이터 전처리

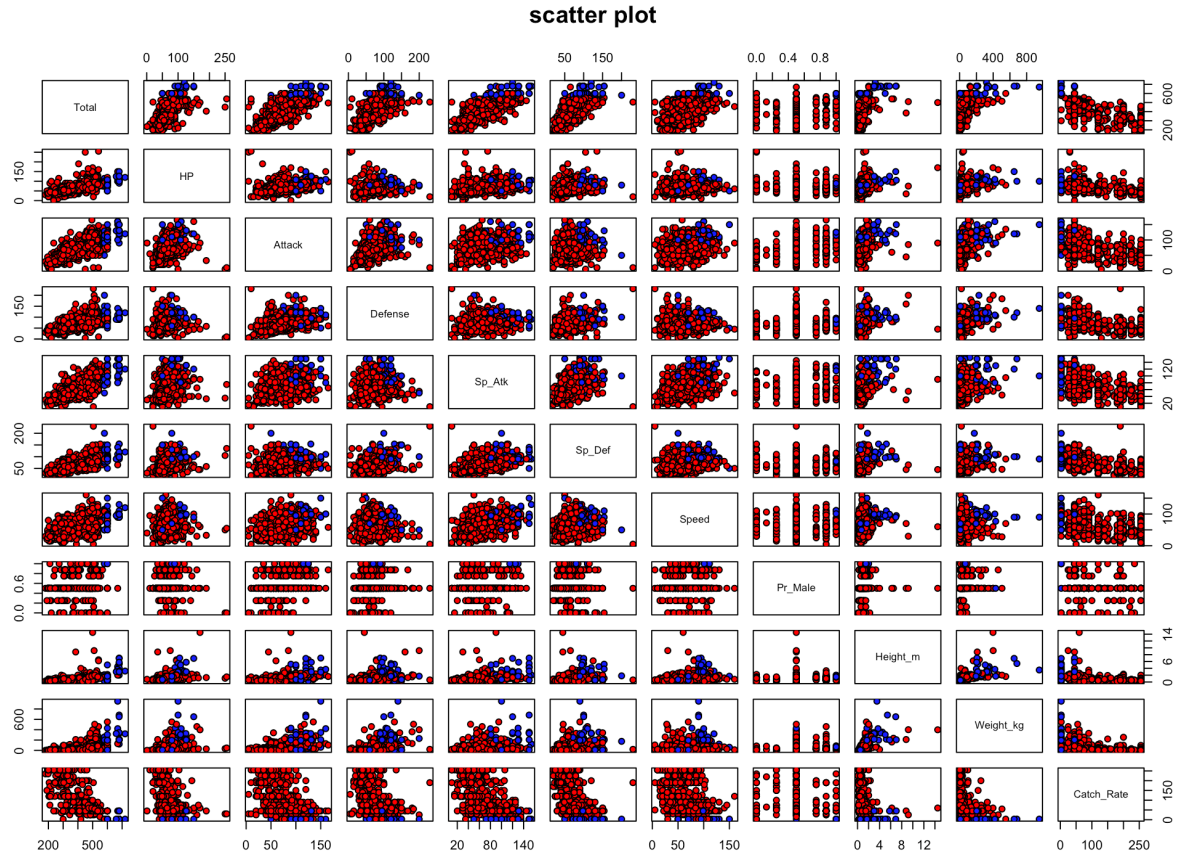
Dimension reduction과 modeling을 하기 위해선 각각의 전제 조건을 만족해야한다. LCM(Linear conditional mean)과 CCV(Constant Conditional Variance with respect to beta) 조건을 만족하기 위해 다음과 같은 전처리를 진행한다.

a. Data selection

우선 Dimension reduction을 진행하기 위해 범주형 변수들과 유니크한 값만을 가지는 포켓몬 도감 번호(Number)를 제거한다. 단 이때, target 변수인 'isLegendary'는 유지한다. 따라서 최종적으로 target 변수인 'isLegendary'와 11개의 연속형 변수를 이용한다.

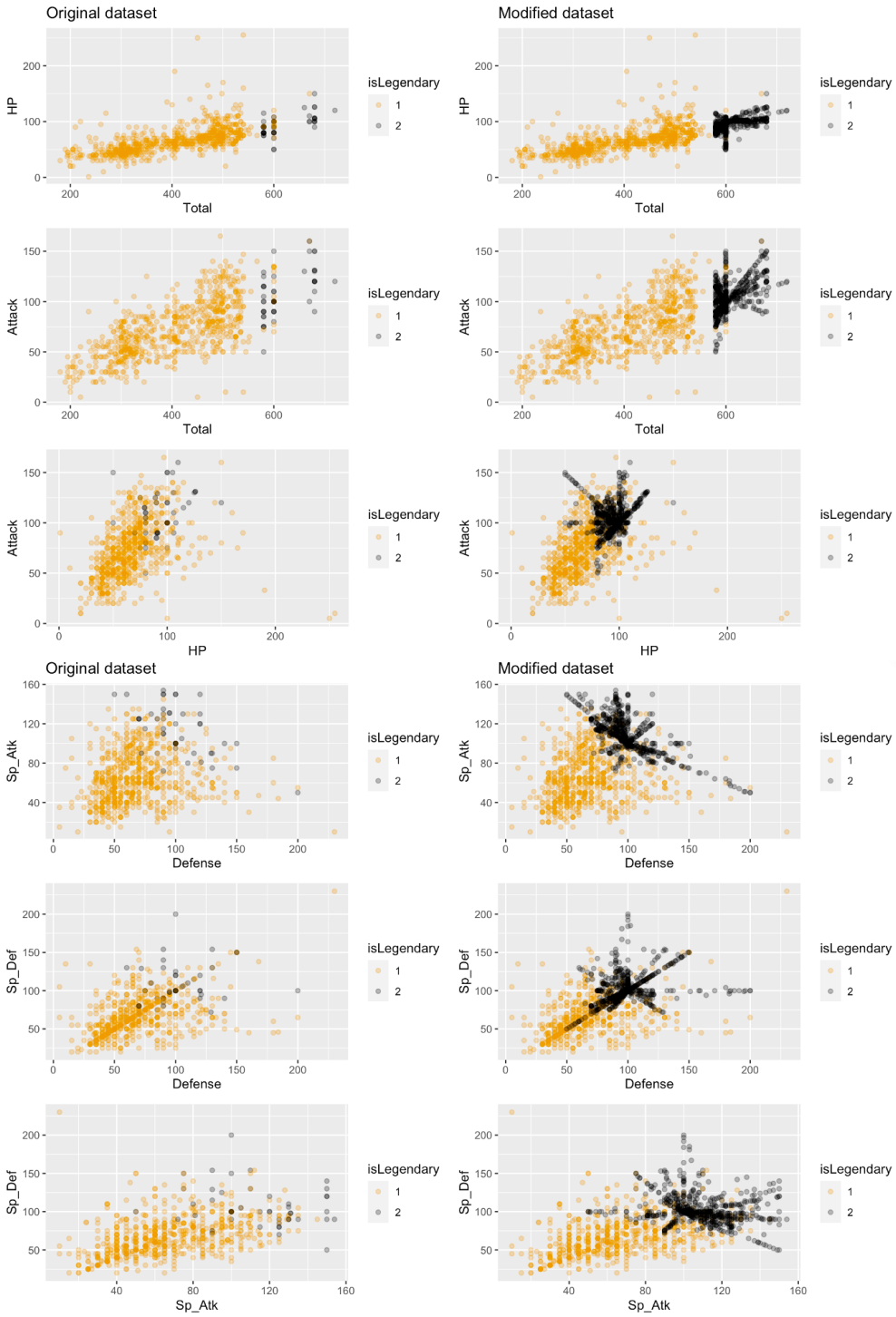
b. Oversampling

최종 데이터에 대해 산점도를 그려보면 [그림 4]와 같다. 앞서 언급했듯, target 변수인 isLegendary에 데이터 불균형이 있는 것을 확인할 수 있다. 이러한 불균형을 유지한 상태로 모델링을 진행 할 시, 모델의 성능은 좋게 나올 수밖에 없는 상황이다. 따라서 Oversampling을 통해 데이터 비율을 맞춘다.



[그림 4. Scatter plot by isLegendary]

오버 샘플링은 SMOTE 알고리즘 기반인 MWMOTE를 이용한다. MWMOTE는 소수의 클러스터에 대해 두 클래스의 경계 근처에 더 높은 가중치를 할당하여 데이터를 생성, resampling하는 방식이다. 오버 샘플링을 진행한 후 데이터 변화는 [그림 5]와 같다. 원 데이터와 샘플링 후 데이터를 비교해보면, 원 데이터 클래스에서 크게 동떨어진 값을 가지는 것으로는 보이지 않으며 각 클래스에 해당하는 observation이 667와 30에서 667와 630으로 전설의 포켓몬 비율이 증가하였다. 즉, 약 1:1의 비율로 두 클래스가 비슷하게 구성되었으며 이를 이용한다.

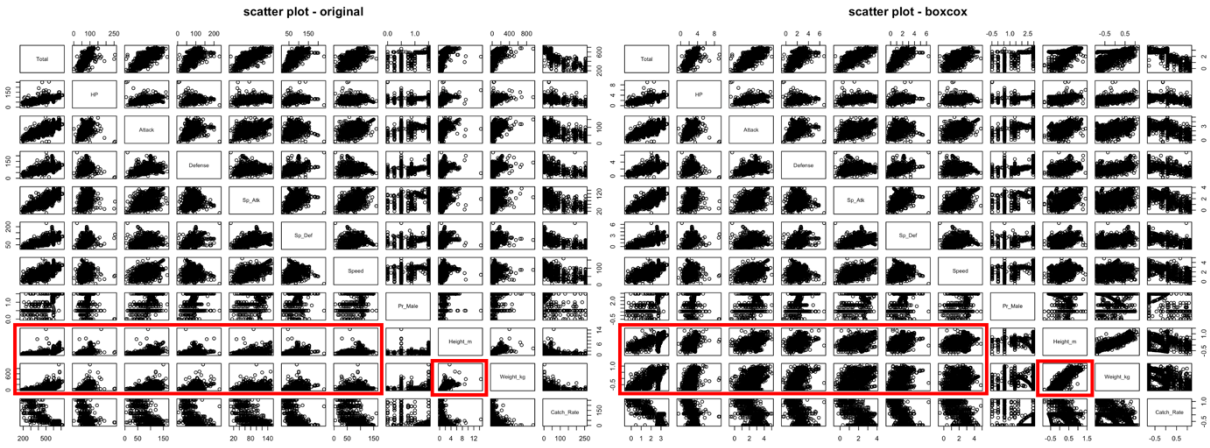


[그림 5. 원 데이터와 오버 샘플링 데이터 비교]

c. Box-Cox Transformation

데이터에 대한 산점도를 그려본 결과 대부분의 변수들이 Elliptical shape의 형태를 보였지만 'Height_m', 'Weight_kg'등의 변수는 Elliptical shape과 멀어 보이는 것을 확인하였다. 따라서 변수들의 모든 값에 음수가 없으므로 Box-Cox Transformation을 진행하였으며, 그 결과는 [그림 6]의 오른쪽 산점도와 같다. 그래프의 빨간 박스를 보면 Transformation전과 후를 비교하였을 때

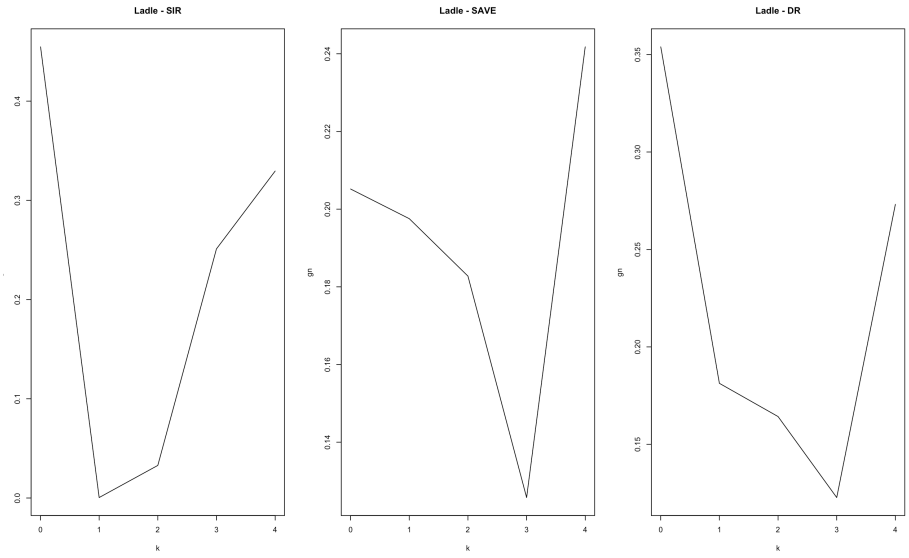
Elliptical shape에 가까워진 것을 볼 수 있다. 따라서, Dimension Reduction과 modeling에 데이터로 이를 이용한다.



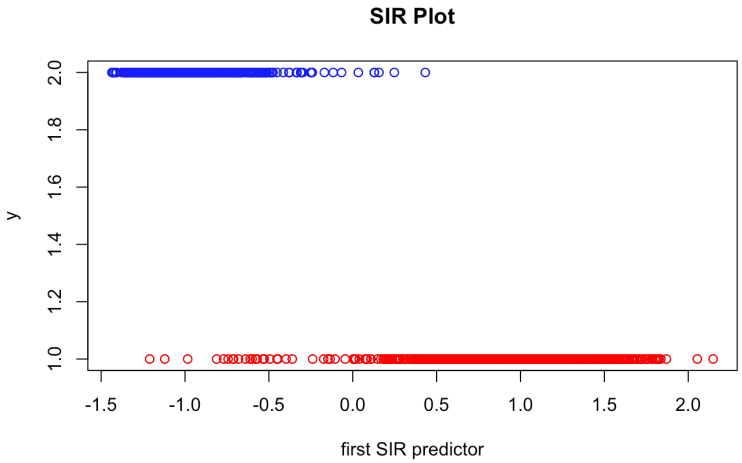
[그림 6. Box-Cox Transformation 진행 전(왼) 산점도와 진행 후(오) 산점도]

4. Dimension Reduction

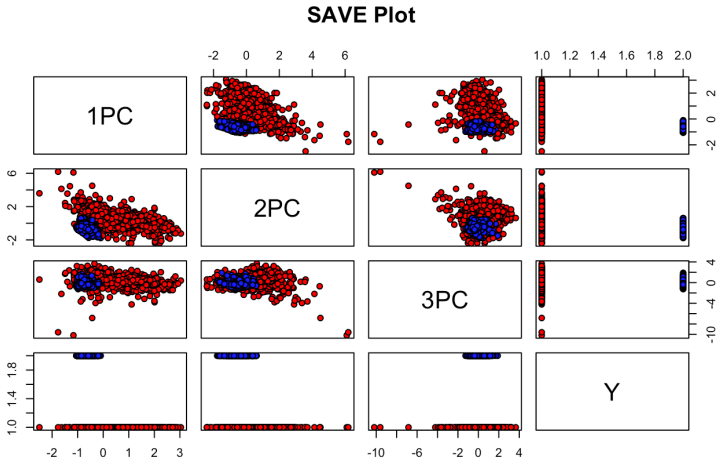
본 프로젝트에서 사용한 차원 축소 방법은 SIR, SAVE, DR로 이를 데이터에 적용하기 전에 Ladle estimator를 이용하여 각 방법에 적절한 차원의 수를 선택하고자 한다. 그 결과는 [그림 7]과 같으며, SIR의 차원은 1, SAVE의 차원은 3, DR의 차원은 3으로 각 방법을 데이터에 적용한다. 전체 데이터에 선택된 차원으로 SIR, SAVE, DR을 적용한 결과는 [그림 8]~[그림 10]과 같다. [그림 8]은 SIR 방식의 차원 축소를 전체 데이터에 적용한 후 y class에 따라 red(isLegendary: No)와 blue(isLegendary: Yes) 색을 통해 그래프를 그려보았다. [그림 9]와 [그림 10]도 동일한 방법으로 그래프를 그린 결과로, 데이터를 잘 설명하는 차원을 한눈에 볼 수 있다.



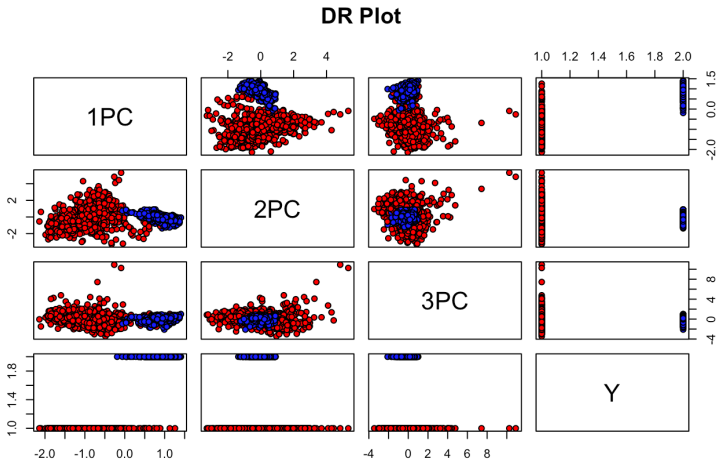
[그림 7. Ladle result SIR / SAVE / DR]



[그림 8. SIR sufficient predictor]



[그림 9. SAVE sufficient predictor]



[그림 9. DR sufficient predictor]

5. Modeling

모델링을 하기 위해 R library인 creditmodel을 이용하여 train과 test 데이터를 7:3의 비율로 나누었다. 즉, train data은 908개로 y class의 비율은 473:435 이고, test data는 389개이며 y class 비율은 194:195 이다. 차원 축소를 적용한 경우와 적용하지 않은 경우 분류 모형의 분류 결과를 비교 분석해 그 성능을 비교했다. 분류 모형으로는 SVM(Support Vector Machine), Random Forest, XGBoost, Naive Bayes을 이용하였다.

Model	X_original	SIR(d=1)	SAVE(d=3)	DR(d=3)
SVM	**0.9922879**	0.966581	**0.9794344**	0.907455
Random Forest	**0.9922879**	0.9640103	0.9717224	**0.9717224**
XGBoost	0.9820051	0.9614396	**0.9794344**	0.9691517
Naive Bayes	0.9794344	**0.9717224**	0.6632391	0.7789203

[표 2. Accuracy Comparison]

머신러닝 분류 모형을 사용했을 때 정확도는 [표 2]와 같다. 원데이터에 분류 모형을 적용한 경우 SVM, Random Forest 모형의 정확도가 약 99%로 가장 높았다. 차원 축소를 적용한 데이터의 경우 SIR은 Naive Bayes 모형이 약 97%, SAVE는 SVM과 XGBoost 모형이 약 97%, DR은 Random Forest 모형이 약 97%로 가장 정확도가 높았다. Naive Bayes 모형의 경우 SIR을 제외한 차원 축소 방법에 대해 약 66%~77%의 정확도를 보이며 제일 낮은 성능을 볼 수 있었다. 또한, Random Forest와 XGBoost는 차원 축소 방법 종류에 상관없이 대체로 약 96~97%로 정확도에 큰 변화가 없었다.
차원 축소 여부에 대한 정확도 비교를 보면, 차원 축소를 적용한 경우의 모형 정확도가 더 낮았다. 따라서, 본 프로젝트를 진행한 포켓몬 데이터에 대해서는 차원 축소를 적용하지 않는 편이 전설의 포켓몬 여부를 분류하는 것에 있어 더 유의미하다는 것을 알 수 있다.

6. 결론

지금까지 차원 축소 방법을 이용해 포켓몬 데이터의 전설의 포켓몬 여부를 분류해보고 차원 축소 적용 전과 후의 성능을 비교해보았다. SVM, Random Forest, XGBoost, Naive Bayes 분류 모형 중 차원 축소 방법인 SIR의 경우 Naive Bayes가 가장 좋은 성능을 보였고, SAVE는 SVM이 가장 좋은 성능을 보였다. 마지막으로 DR은 XGBoost가 가장 정확도가 높은 것을 확인했다. 하지만 모든 차원 축소 방법 보다 차원 축소를 진행하지 않은 데이터에 대해 모형을 적용한 것이 대부분의 모형에서 약 99%의 정확도로 가장 높은 성능을 보였다. 따라서 포켓몬 데이터를 이용하여 전설의 포켓몬 여부를 분류하는 태스크에 대해 차원 축소를 이용하는 것은 모델의 성능을 저하하는 요소가

될 수 있음을 설명한다.

본 프로젝트의 한계 및 의의는 다음과 같다. 먼저 데이터의 크기가 작다는 점이다. 프로젝트에 사용된 데이터는 6세대까지의 포켓몬 데이터로 총 721마리의 포켓몬을 통해 분석을 진행한 것이다. 현재 9세대까지 포켓몬 시리즈가 확장되었고, 새롭게 추가된 포켓몬을 포함하여 1000종이 넘는 포켓몬 데이터를 사용할 수 있었다면 더 다양한 결과를 볼 수 있었을 것이라 생각한다. 그럼에도 주어진 포켓몬의 정보를 통해 전설의 포켓몬을 분류 할 시 차원 축소를 이용하는 것은 긍정적인 결과를 기대하기 힘들다는 것을 알 수 있었으며, 종족값 및 여러 정보를 통해 포켓몬을 새롭게 생성하면 게임 내에서 어느 위치에 해당하는지 등 몬스터 레벨 스케일링에 도움이 될 수 있다는 의의가 있다고 생각한다.