true

Prediction of Problematic Internet Use

Group5(廖芷萱、詹雅鈞、李姿慧、謝沛恩、李敏榕)

Table of contents

資料介紹

資料來源為 Kaggle 競賽提供的 Healthy Brain Network (HBN) 資料集

其為一臨床樣本，包含3960名年齡介於 5 至 22 歲的青少年

資料集中包含以下兩類元素：

- 體能活動資料
- 網路使用行為資料

目標與動機

基於兒童和青少年的體能活動、身體測量、心理健康及網路行為等特徵，建立成癮嚴重程度(sii)的預測模型

資料描述

訓練資料集包含3960筆樣本，共計82個變數。其中包含59個解釋變數，分為以下類別：

- Demographics
- Children’s Global Assessment Scale(兒童全球評估量表)
- Physical Measures
- FitnessGram Vitals and Treadmill(健體測驗生命指標及跑步機測試)
- FitnessGram Child(兒童版健體測驗)

資料描述

- Bio-electric Impedance Analysis(生物電阻抗分析)
- Physical Activity Questionnaire (Adolescents)
- Physical Activity Questionnaire (Children)
- Sleep Disturbance Scale(兒童睡眠障礙量表)
- Internet Use

資料描述

研究之反應變數為網絡成癮嚴重程度 (Severity Impairment Index, SII)

該指數依據總分範圍將樣本分為四個層級：

- 0 = 無 (None)
- 1 = 輕度 (Mild)
- 2 = 中度 (Moderate)
- 3 = 重度 (Severe)

其定義基於Parent-Child InternetAddictionTest(PCIAT)，並以PCIAT_Total量化嚴重程度

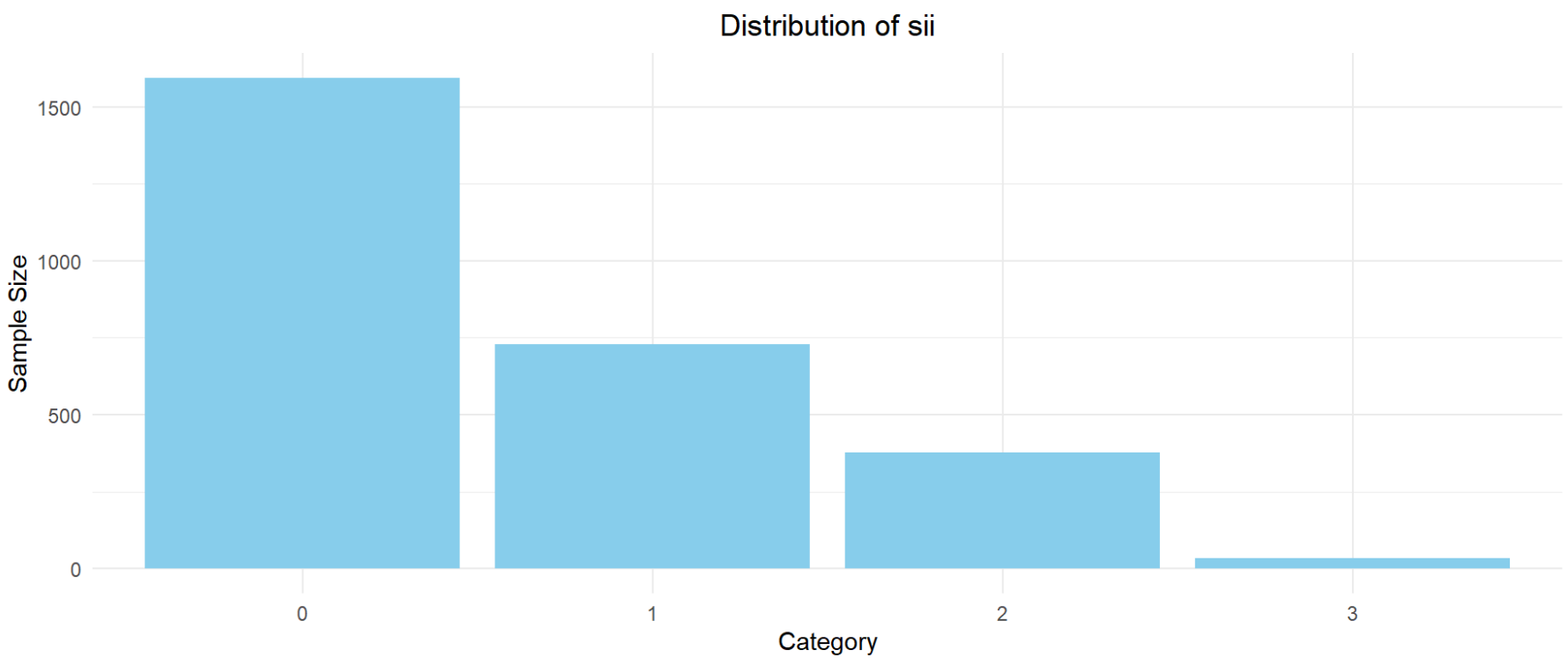
遺失值

檢視資料中反應變數與解釋變數的缺失值情況

- Code

遺失值

- Code


Distribution of sii
Sample Size
1500
1000
500
0
0
1
2
3
Category

反應變數分析

反應變數-網絡成癮嚴重程度的分佈顯示出類別不平衡的現象

此類別不平衡可能會對後續分析或模型訓練過程中的結果產生偏差

將類別 2（Moderate）與類別 3（Severe）合併為類別 2（Moderate-Severe）

可有助於提升模型的穩定性，並減少少數類別樣本數量對結果的過度影響

遺失值

```r
train$sii <- ifelse(train$sii == 3, 2, train$sii)

train$sii <- factor(train$sii, levels = c(0,1, 2) ,ordered = TRUE)

table(train$sii) # 檢查合併後的類別分布
```

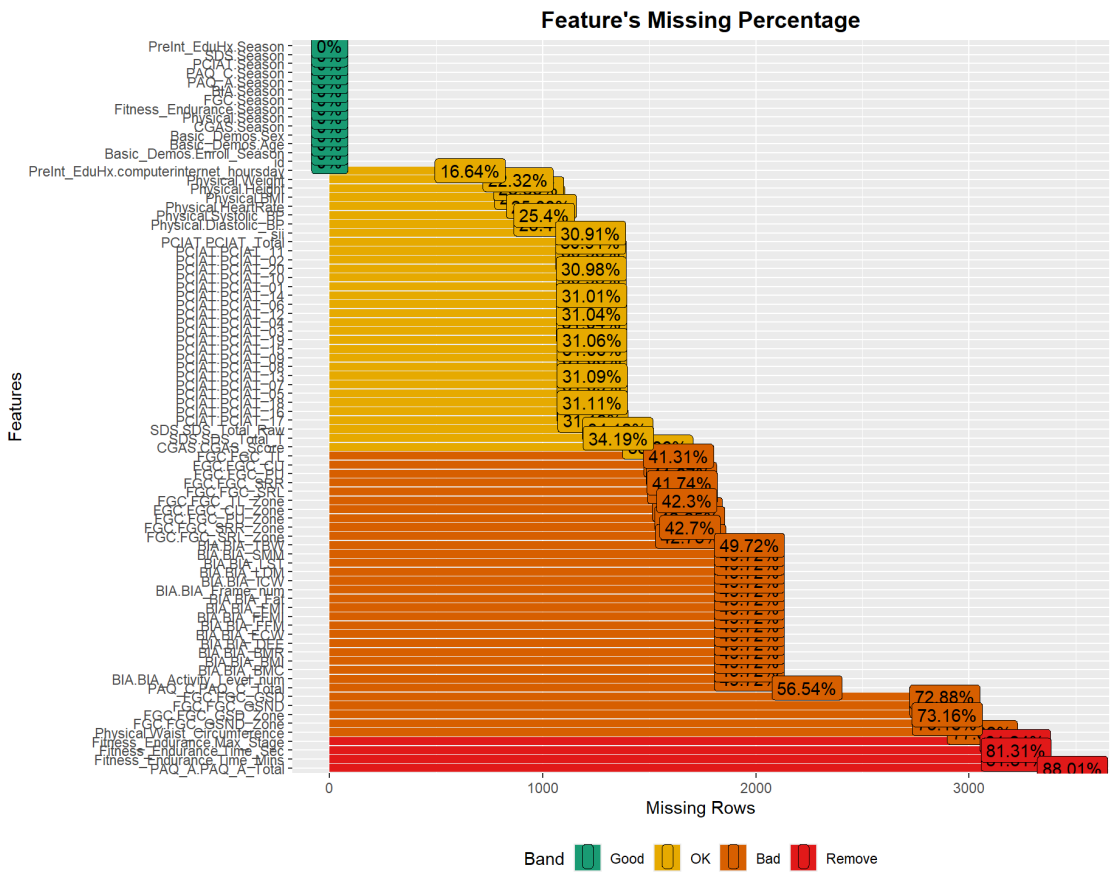
```
   0    1    2 
1594  730  412 
```

遺失値

- Code

遺失值

- Code


Feature's Missing Percentage
Features
PreInt_EduHx.Season
SDS.Season
PCIAT.Season
PAQ_C.Season
PAQ_A.Season
BIA.Season
FGC.Season
Fitness_Endurance.Season
Physical.Season
CGAS.Season
Basic_Demos.Sex
Basic_Demos.Age
Basic_Demos.Enroll_Season
id
PreInt_EduHx.computerinternet_hoursday
Physical.Weight
Physical.Height
Physical.BMI
Physical.HeartRate
Physical.Systolic_BP
Physical.Diastolic_BP
Sii
PCIAT.PCIAT_Total
PCIAT.PCIAT_01
PCIAT.PCIAT_02
PCIAT.PCIAT_20
PCIAT.PCIAT_10
PCIAT.PCIAT_11
PCIAT.PCIAT_14
PCIAT.PCIAT_06
PCIAT.PCIAT_12
PCIAT.PCIAT_04
PCIAT.PCIAT_03
PCIAT.PCIAT_19
PCIAT.PCIAT_17
PCIAT.PCIAT_09
PCIAT.PCIAT_08
PCIAT.PCIAT_13
PCIAT.PCIAT_07
PCIAT.PCIAT_05
PCIAT.PCIAT_18
PCIAT.PCIAT_16
PCIAT.PCIAT_15
SDS.SDS_Total_Raw
SDS.SDS_Total_T
CGAS.CGAS_Score
FGC.FGC_TL
FGC.FGC_CU
FGC.FGC_PR
FGC.FGC_SRR
FGC.FGC_SRL
FGC.FGC_TL_Zone
FGC.FGC_CU_Zone
FGC.FGC_PR_Zone
FGC.FGC_SRR_Zone
FGC.FGC_SRL_Zone
BIA.BIA_FMI
BIA.BIA_SMM
BIA.BIA_LST
BIA.BIA_TBW
BIA.BIA_ICW
BIA.BIA_Frame_num
BIA.BIA_LDM
BIA.BIA_FFMI
BIA.BIA_FFM
BIA.BIA_ECW
BIA.BIA_BFP
BIA.BIA_BMR
BIA.BIA_BMI
BIA.BIA_BMC
BIA.BIA_Activity_Level_num
PAQ_C.PAQ_C_Total
FGC.FGC_CU_Zone
FGC.FGC_GSND
FGC.FGC_GSND_Zone
FGC.FGC_GSND_Zone
PhysicalWaist_Circumference
Fitness_Endurance.Max_Stage
Fitness_Endurance.Time_Sec
Fitness_Endurance.Time_Mins
PAQ_A.PAQ_A_Total
0%
16.64%
25.4%
30.91%
30.98%
31.01%
31.04%
31.06%
31.09%
31.11%
34.19%
41.31%
41.74%
42.3%
42.7%
49.72%
56.54%
72.88%
73.16%
81.31%
0
1000
2000
3000
Missing Rows
Band
Good
OK
Bad
Remove

遺失值分析

資料集中存在多個具有較高比例遺失值的變數

將根據變數的含義及其與其他變數的相關性進行變數選擇

變數選擇初步分析

檢查Demographics，結果如下

▶ Code

Field	missing %	Description
Basic_Demos-Enroll_Season	0	Season of enrollment
Basic_Demos-Age	0	Age of participant
Basic_Demos-Sex	0	Sex of participant

變數選擇初步分析

檢查Children’s Global Assessment Scale，結果如下

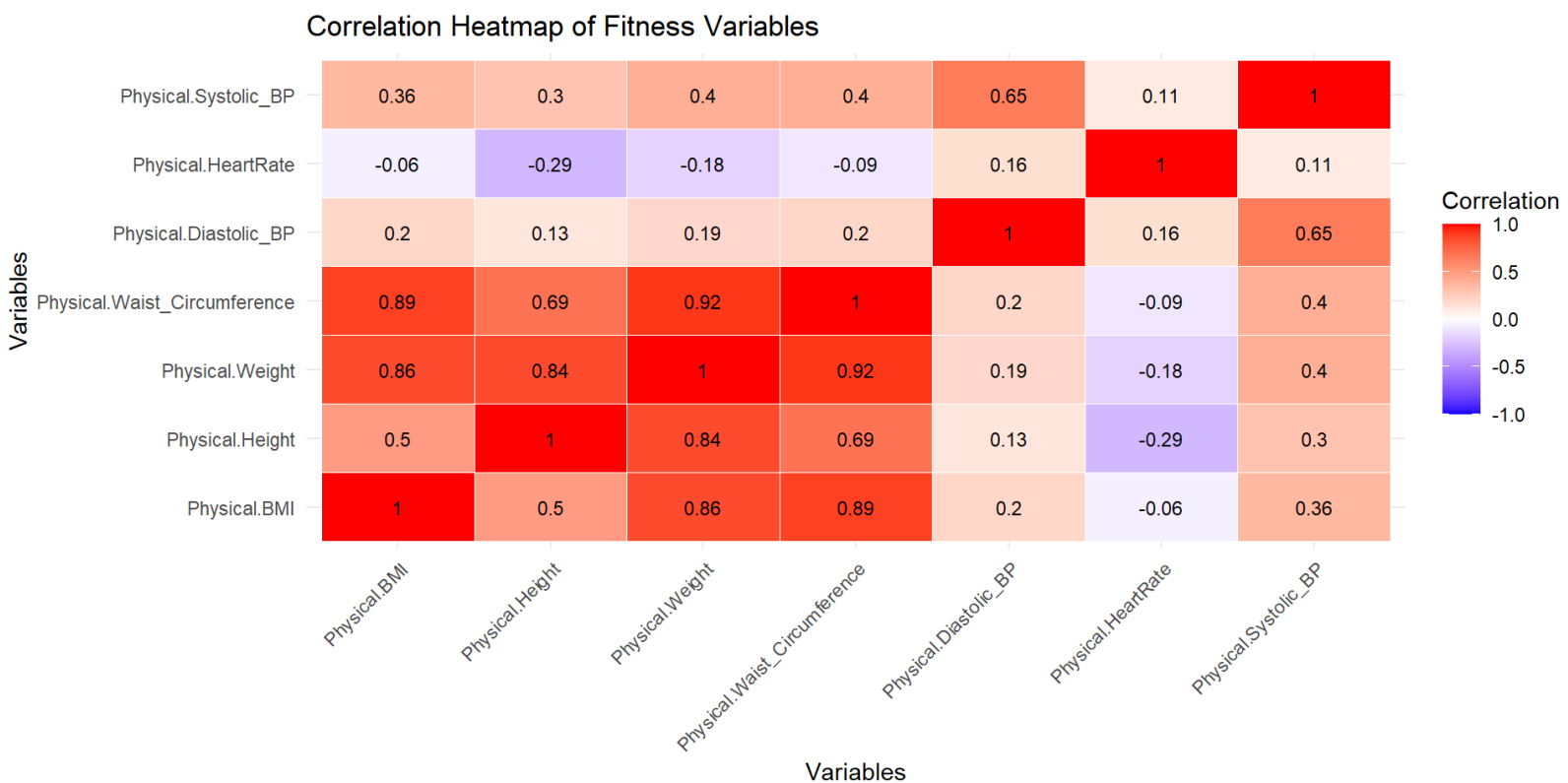
▶ Code

Field	missing %	Description
CGAS-Season	0.00	Season of participation
CGAS-CGAS_Score	38.86	Children's Global Assessment Scale Score

變數選擇初步分析

檢查Physical Measures之相關性，使用相關係數矩陣分析

- Code



變數選擇初步分析

結果如下

▶ Code

Field	missing %	Description
Physical-Season	0.00	Season of participation
Physical-BMI	**23.69**	**Body Mass Index (kg/m^2)**
Physical-Height	**23.56**	**Height (in)**
Physical-Weight	**22.32**	**Weight (lbs)**
Physical-Waist_Circumference	**77.32**	**Waist circumference (in)**
Physical-Diastolic_BP	25.40	Diastolic BP (mmHg)
Physical-HeartRate	25.08	Heart rate (beats/min)
Physical-Systolic_BP	25.40	Systolic BP (mmHg)

變數選擇初步分析

檢查FitnessGram Vitals and Treadmill

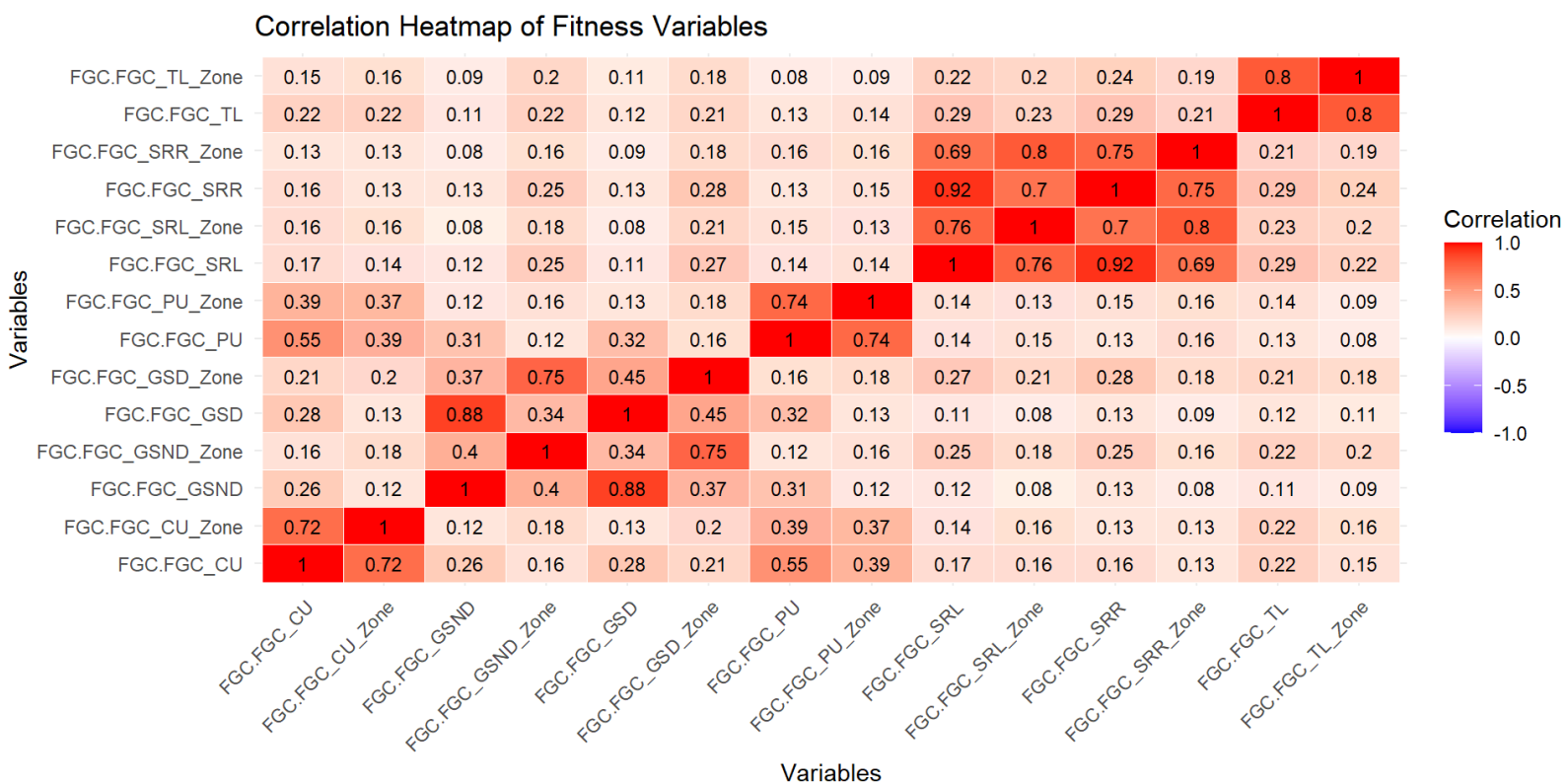
▶ Code

Field	missing %	Description
Fitness_Endurance-Season	0.00	Season of participation
Fitness_Endurance-Max_Stage	81.24	Maximum stage reached
Fitness_Endurance-Time_Mins	81.31	Exact time completed: Minutes
Fitness_Endurance-Time_Sec	81.31	Exact time completed: Seconds

變數選擇初步分析

檢查FitnessGram Child之相關性，使用相關係數矩陣分析

- Code



變數選擇初步分析

結果如下

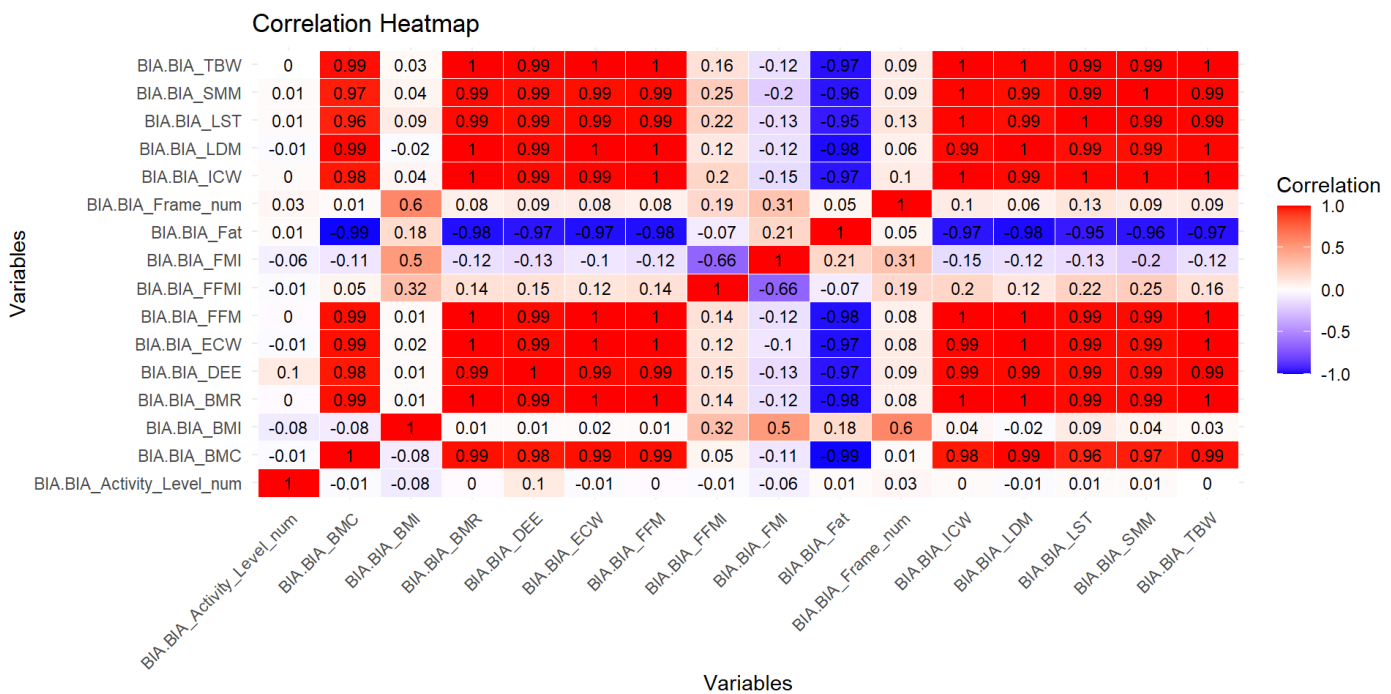
▶ Code

Field	missing %	Description
FGC-Season	0.00	Season of participation
FGC-FGC_CU	41.36	Curl up total
FGC-FGC_CU_Zone	42.37	Curl up fitness zone
FGC-FGC_GSND	72.88	Grip Strength total (non-dominant)
FGC-FGC_GSND_Zone	73.18	Grip Strength fitness zone (non-dominant)
FGC-FGC_GSD	72.88	Grip Strength total (dominant)
FGC-FGC_GSD_Zone	73.16	Grip Strength fitness zone (dominant)
FGC-FGC_PU	41.67	Push-up total
FGC-FGC_PU_Zone	42.65	Push-up fitness zone
FGC-FGC_SRL	**41.79**	**Sit & Reach total (left side)**
FGC-FGC_SRL_Zone	**42.75**	**Sit & Reach fitness zone (left side)**
FGC-FGC_SRR	**41.74**	**Sit & Reach total (right side)**
FGC-FGC_SRR_Zone	**42.70**	**Sit & Reach fitness zone (right side)**
FGC-FGC_TL	41.31	Trunk lift total
FGC-FGC_TL_Zone	42.30	Trunk lift fitness zone

變數選擇初步分析

檢查Bio-electric Impedance Analysis之相關性，使用相關係數矩陣分析

- Code



變數選擇初步分析

結果如下

▶ Code

Field	missing %	Description
BIA-Season	0.00	Season of participation
BIA-BIA_Activity_Level_num	49.72	Activity Level
BIA-BIA_BMC	**49.72**	**Bone Mineral Content**
BIA-BIA_BMI	49.72	Body Mass Index
BIA-BIA_BMR	**49.72**	**Basal Metabolic Rate**
BIA-BIA_DEE	**49.72**	**Daily Energy Expenditure**
BIA-BIA_ECW	**49.72**	**Extracellular Water**
BIA-BIA_FFM	**49.72**	**Fat Free Mass**
BIA-BIA_FFMI	49.72	Fat Free Mass Index
BIA-BIA_FMI	49.72	Fat Mass Index
BIA-BIA_Fat	49.72	Body Fat Percentage
BIA-BIA_Frame_num	49.72	Body Frame
BIA-BIA_ICW	**49.72**	**Intracellular Water**
BIA-BIA_LDM	**49.72**	**Lean Dry Mass**
BIA-BIA_LST	**49.72**	**Lean Soft Tissue**
BIA-BIA_SMM	**49.72**	**Skeletal Muscle Mass**
BIA-BIA_TBW	**49.72**	**Total Body Water**

變數選擇初步分析

檢查Physical Activity Questionnaire-Adolescents

▶ Code

Field	missing %	Description
PAQ_A-Season	0.00	Season of participation
PAQ_A-PAQ_A_Total	88.01	Activity Summary Score (Adolescents)

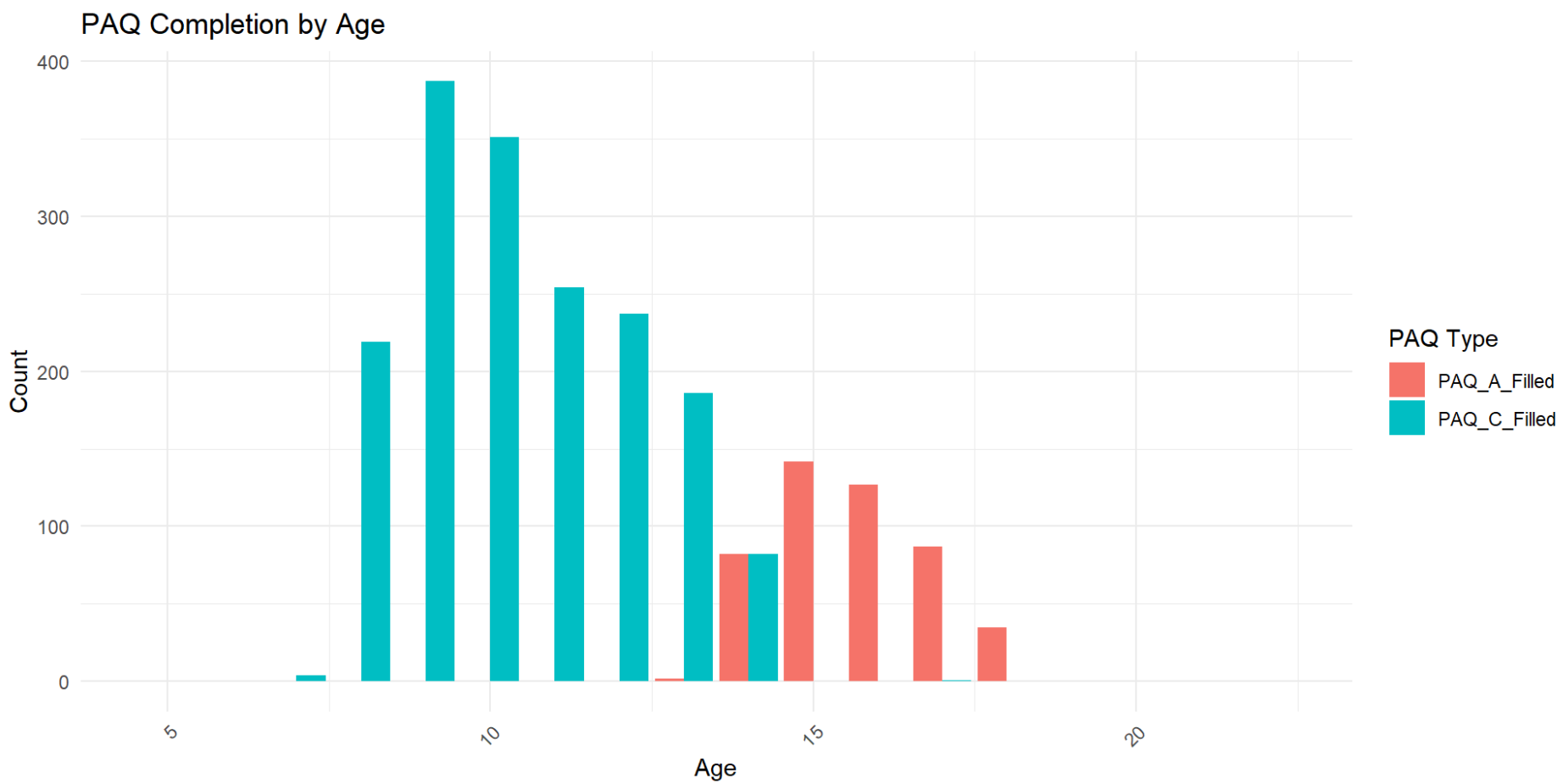
檢查Physical Activity Questionnaire-Children

▶ Code

Field	missing %	Description
PAQ_C-Season	0.00	Season of participation
PAQ_C-PAQ_C_Total	56.54	Activity Summary Score (Children)

統計 PAQ_A 和 PAQ_C

- Code



合併PAQ_A &PAQ_C

- Code

身體活動問卷青少年(適用於14-19歲)與兒童版(適用於8-14歲)兩者數據互斥，將其合併為一個變數

若數據同時來自兩個測驗，則取平均值

變數選擇初步分析

檢查Sleep Disturbance Scale

▶ Code

Field	missing %	Description
SDS-Season	0.00	Season of participation
SDS-SDS_Total_Raw	34.12	Total Raw Score
SDS-SDS_Total_T	34.19	Total T-Score

變數選擇初步分析

檢查Internet Use

- Code

Field	missing %	Description
PreInt_EduHx-Season	0.00	Season of participation
PreInt_EduHx-computerinternet_hoursday	16.64	Hours of using computer/internet

最終變數選取

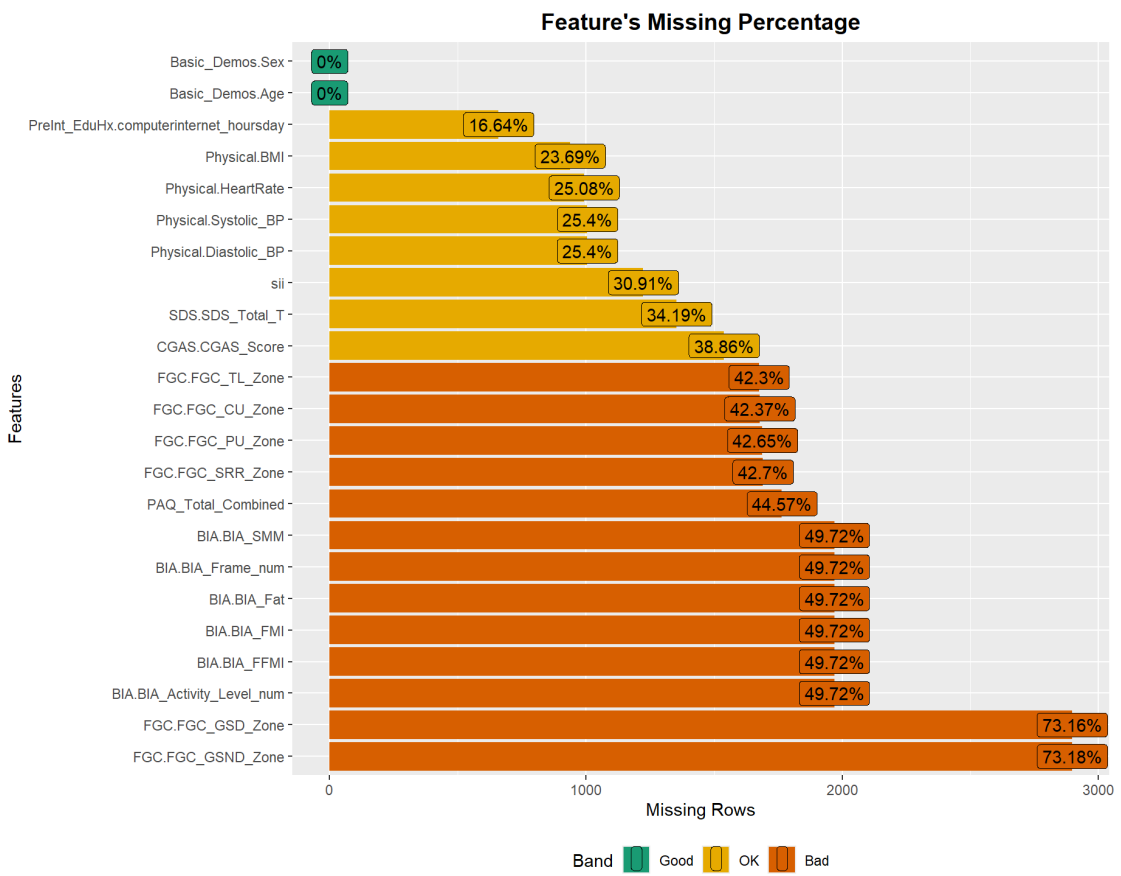
```
# 根據變數含義選取變數
selected<-c("Basic_Demos.Age", "Basic_Demos.Sex", "CGAS.CGAS_Score", "Physi
  "Physical.HeartRate", "Physical.Systolic_BP", "FGC.FGC_CU_Zone", "FGC.FGC
  "FGC.FGC_GSD_Zone", "FGC.FGC_PU_Zone", "FGC.FGC_SRR_Zone", "FGC.FGC_TL_Zo
  "BIA.BIA_Activity_Level_num", "BIA.BIA_FFMI", "BIA.BIA_FMI", "BIA.BIA_Fat
  "BIA.BIA_Frame_num", "BIA.BIA_SMM", "PAQ_Total_Combined",
  "SDS.SDS_Total_T", "PreInt_EduHx.computerinternet_hoursday", "sii")
train_1<- train[,selected]
```

最終變數遺失值分析

```r
miss_lot <- plot_missing(train_1) + #原始資料遺失值
  ggtitle("Feature's Missing Percentage") +  # 新增標題
  theme(
    plot.title = element_text(hjust = 0.5, size = 14, face = "bold"),  # 標
    plot.margin = margin(20, 20, 20, 20)  # 增加下方的邊距
  )
```

最終變數遺失值分析

- Code



最終變數遺失值分析

▶ Code

FGC.FGC_GSND_Zone和FGC.FGC_GSD_Zone的缺失值比例均超過70%，將其刪除

▶ Code

對有反應變數-網絡成癮嚴重程度（SeverityImpairment Index,SII）之缺失值的資料刪除，避免遺失值對分析結果的影響

插補缺失值

對於缺失值比例低於50%的變數，採用了多重插補方法（Multiple Imputation by Chained Equations, MICE）

多重插補的迭代次數（iteration）為 50 次，並生成了 5 個插補後的資料集

▶ Code

模型訓練

Ordinal Logistic Regression

有序邏輯斯迴歸用來處理反應變數為順序類別變數的資料

$$logit(P(Y \leq j)) = \log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) = \alpha_j - x^\top \beta,\ j = 1, 2, ..$$

Ordinal Logistic Regression

- Code

特性	Logistic.Regression	Ordinal.Logistic.Regression
資料型態	適用於二元類別資料	適用於有序類別資料
類別數量	二元類別	多個有序類別
類別順序考量	忽略類別之間的順序關係	考慮類別之間的順序
模型假設	預測的 log-odds 為線性函數的形式	假設平行線和反應變數為順序變數
解釋重點	解釋單一類別相對於另一類別的機率 (Odds Ratio)	解釋類別累積機率，或在不同閾值間的隱變數變化
模型輸出	每個觀測值歸屬於某一類別的機率	預測每個觀測值落在某一類別或以上的累積機率
適用情境	適用於二元分類問題	適用於有序類別問題
效能表現	快速、適合處理大量二元分類問題	計算較複雜，適合處理類別數較多且有序的問題
實現方式	R 套件glm 或 Python 套件statsmodels	R 套件MASS::polr 或 Python套件 statsmodels 的 OrderedModel

Ordinal Logistic Regression

```r
# 建模
fit <- with(mice_train,polr(sii ~ Basic_Demos.Age + Basic_Demos.Sex + CGAS.
, Hess = TRUE))
library(mice)
pooled <- pool(fit)
```

有些變數的fmi還是有點偏高(>0.5)

如Physical.BMI、FGC.FGC_PU_Zone、FGC.FGC_TL_Zone、BIA.BIA_FFMI、BIA.BIA_FMI、BIA.BIA_Frame_num2

Ordinal Logistic Regression

```
1 #summary(pooled)
```

模型結果以estimate、std.error、statistic、P-value呈現

如變數在顯著水準為0.05下，為此模型的顯著變數

表示這些變數可能對網絡成癮嚴重程度(sii)有影響：

Basic_Demos.Age、Physical.HeartRate、
BIA.BIA_Frame_num2、SDS.SDS_Total_T、
PreInt_EduHx.computerinternet_hoursday

Ordinal Logistic Regression

- Code

- Code

	x
Accuracy	0.6208791
Kappa	0.3571736

準確率為 62.08%

QWK值為 0.357

表示模型對有序分類的預測有一定效果

Ordinal Forest

一種隨機森林演算法的變體，專門用來處理有序類別變數

將數據中的類別（例如：低、中、高）視為具有順序的數值

Ordinal Forest

工作原理:

- 建立分數集
- 生成回歸森林
- 評估森林效能
- 選擇最佳森林和建立優化的分數集
- 訓練最終的回歸森林

嘗試許多不同的分數集，選擇在預測原始序數反應變數方面表現最佳的分數集

以迭代的方式找到最佳的連續表示法

Ordinal Forest

- Code

特性	普通隨機森林..Random.Forest.	有序森林..Ordinal.Forest.
資料型態	適用於類別型 (分類) 或數值型 (迴歸) 資料	專為處理有序類別資料設計
類別順序考量	忽略類別之間的順序關係	考慮類別之間的順序關係，避免預測結果與真實值相差過遠
分裂準則	以Information Gain或Gini Index為基準	使用順序敏感的分裂準則，優化有序類別的預測
適用情境	適用於所有類別型問題	適用於有序類別問題
模型輸出	類別標籤或數值預測	類別標籤，並確保輸出順序的合理性
誤差懲罰	預測錯誤時，無法區分「小錯誤」與「大錯誤」	預測錯誤時，較大懲罰遠離真實值的錯誤
特徵重要性	提供變數重要性評估，例如基於分裂次數	提供有序資料的變數重要性評估
對類別不平衡的處理	支援權重調整或重新取樣(Resampling)	同樣支援權重調整或重新取樣，並針對小樣本類別提供改進
效能表現	快速、靈活，適合大規模資料	效能較高，但計算量稍多於普通隨機森林
實現方式	R套件randomForest或Python套件scikit-learn	R套件ordfor

Ordinal Forest

- Code

Ordinal Forest

參數設置

- classweights

用來為每個類別賦予不同的權重

Ordinal Forest

參數設置

- perffunction-設定perffunction為”proportional”

在 Ordinal Forest 演算法中，使用gclprop效能函數來評估每個森林的效能

Ordinal Forest

參數設置

gclprop 函數的公式：

- 表示屬於類別 的樣本數量
- 表示總樣本數量
- 表示類別 的 Youden 指數，用於衡量類別 的預測準確度

Ordinal Forest

參數設置

使用 gclprop函數時，模型會傾向於將樣本預測到樣本數量較多的類別

這也意味著較小類別的預測準確度可能會受到影響

Ordinal Forest

▶ Code

	x
Accuracy	0.6051282
Kappa	0.3208794

- 準確率 (Accuracy)：60.51%

模型對目標變數的預測中有超過一半是正確的，但可能不足以滿足高準確性

- QWK值：0.329

模型對有序分類的預測有一定效果，但一致性並不高

CatBoost

CatBoost 是一種基於梯度提升 (Gradient Boosting)的機器學習方法，專為處理分類特徵而設計

此方法提出了有序目標編碼（Ordered Target Encoding）的來避免資料洩漏(data leakage)

CatBoost

Ordered Target Encoding

- 根據樣本的順序對數據進行排序，確保每個樣本的編碼只會參考之前的樣本，而不會使用未來樣本的目標變數資訊
- 計算它與之前所有相同類別值的目標變數的加權平均值
- 平滑參數（ ）減少少數樣本對編碼值的影響，從而避免過度擬合

CatBoost

Ordered Target Encoding

- 表示第 個樣本在類別特徵 上的取值
- 表示第 個樣本在類別特徵 上的取值
- 是指示函數，當條件成立時其值為 1，否則為0
- 是第 個樣本的目標變數值
- 是平滑參數；是所有樣本目標變數的平均值

CatBoost

- Code

方法	優點	缺點	適用情況
OrderedTarget Encoding	防止資訊洩漏（Target Leakage），提高泛化能力	計算較為複雜；需要大量數據支持	類別特徵較多、數據量大
TargetEncoding	簡單、高效率	容易出現資訊洩漏、需處理極端值與樣本不均的問題	類別變數較少
One-HotEncoding	易於理解、編碼時無需計算	類別數量過多時，會導致特徵維度爆炸，增加計算量	類別數量少，特徵較簡單

CatBoost模型建構

- 為增強穩健性，通過**多次隨機排列**和**貪婪選擇特徵組合**來提升預測能力
- 使用Oblivious Tree結構，減少過擬合風險並提高穩定性
- 為解決資料不平衡問題，透過計算每個類別的標記次數來分配權重

CatBoost模型建構

- Code

- Code

CatBoost

▶ Code

Evaluation	Average
Accuracy	0.5744
Quadratic Weighted Kappa	0.2645

平均準確率為 57.44%，準確率較低

平均QWK值為0.2645，顯示模型在處理有序分類時一致性較差，未能充分考慮類別間的順序關係

模型可能需要調整學習率、迭代次數和樹的深度等參數，以提升準確性和穩定性

CatBoost

- Code

CatBoost

- Code

```
Confusion Matrix on Test Set:
```

- Code

```
   test_predictions
      0   1   2
0 213  74  31
1  47  56  43
2  10  23  49
```

從表現最佳模型的混淆矩陣來看，模型在類別0和類別2之間的誤分類較為嚴重

類別0常被誤分為類別2，類別1則常被誤分為類別2

顯示類別1和類別2區別模糊，可能由特徵重疊引起

結論

▶ Code

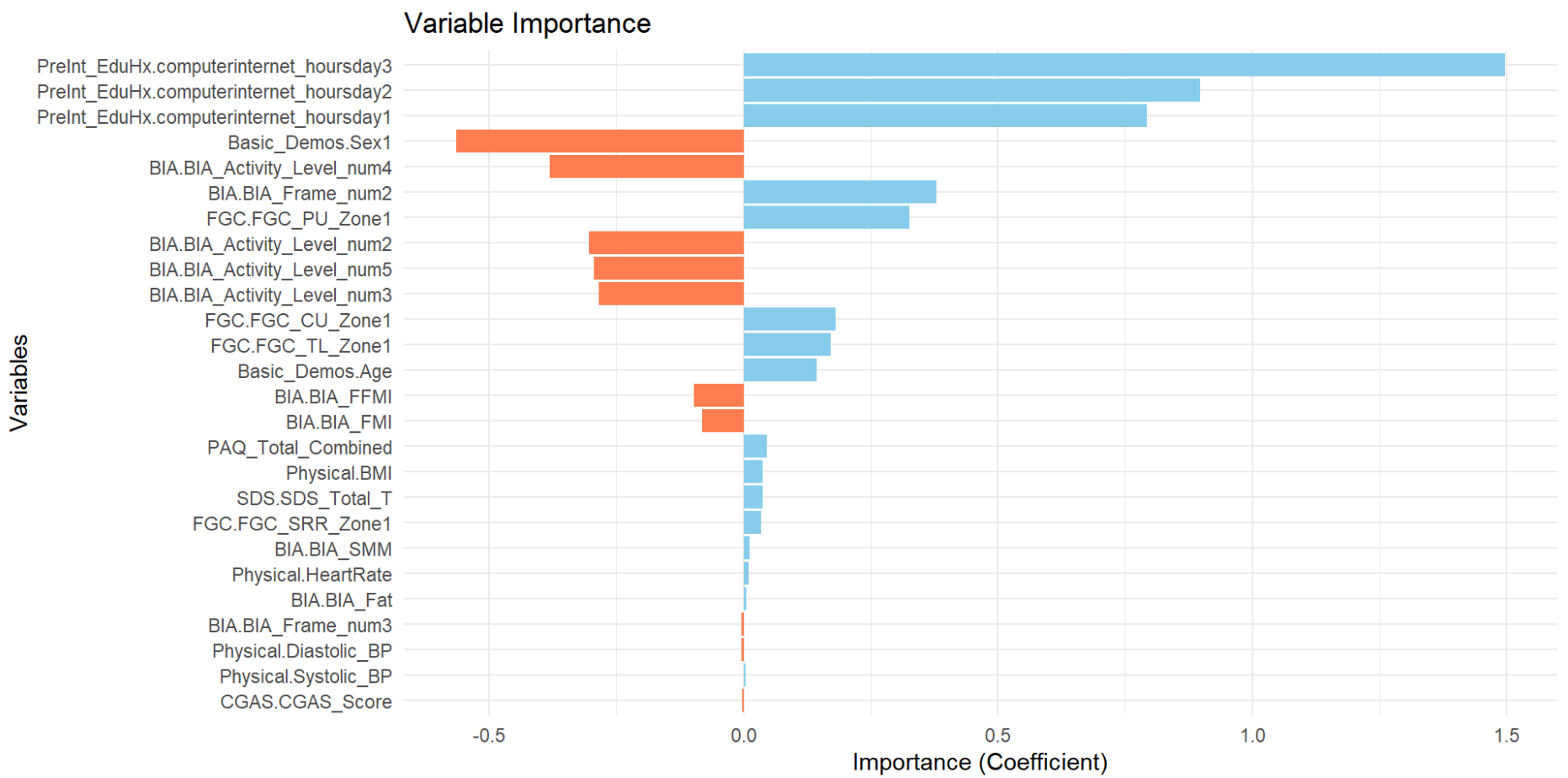
Model	Kappa	Accuracy
Ordinal Logistic Regression	0.3571736	0.6208791
Ordinal Forest	0.3208794	0.6051282
CatBoost	0.2645000	0.5744000

Ordinal Logistic Regression在預測準確性上表現較好，Ordinal Forest表現略遜色於 Ordinal Logistic Regression，但仍能提供相對穩定的預測結果，CatBoost表現較其他兩個模型弱

總體來看，Ordinal Logistic Regression在此次序行類別資料中表現最好

特徵重要性-Ordinal Logistic Regerssion

- Code



特徵重要性-Ordinal Logistic Regerssion

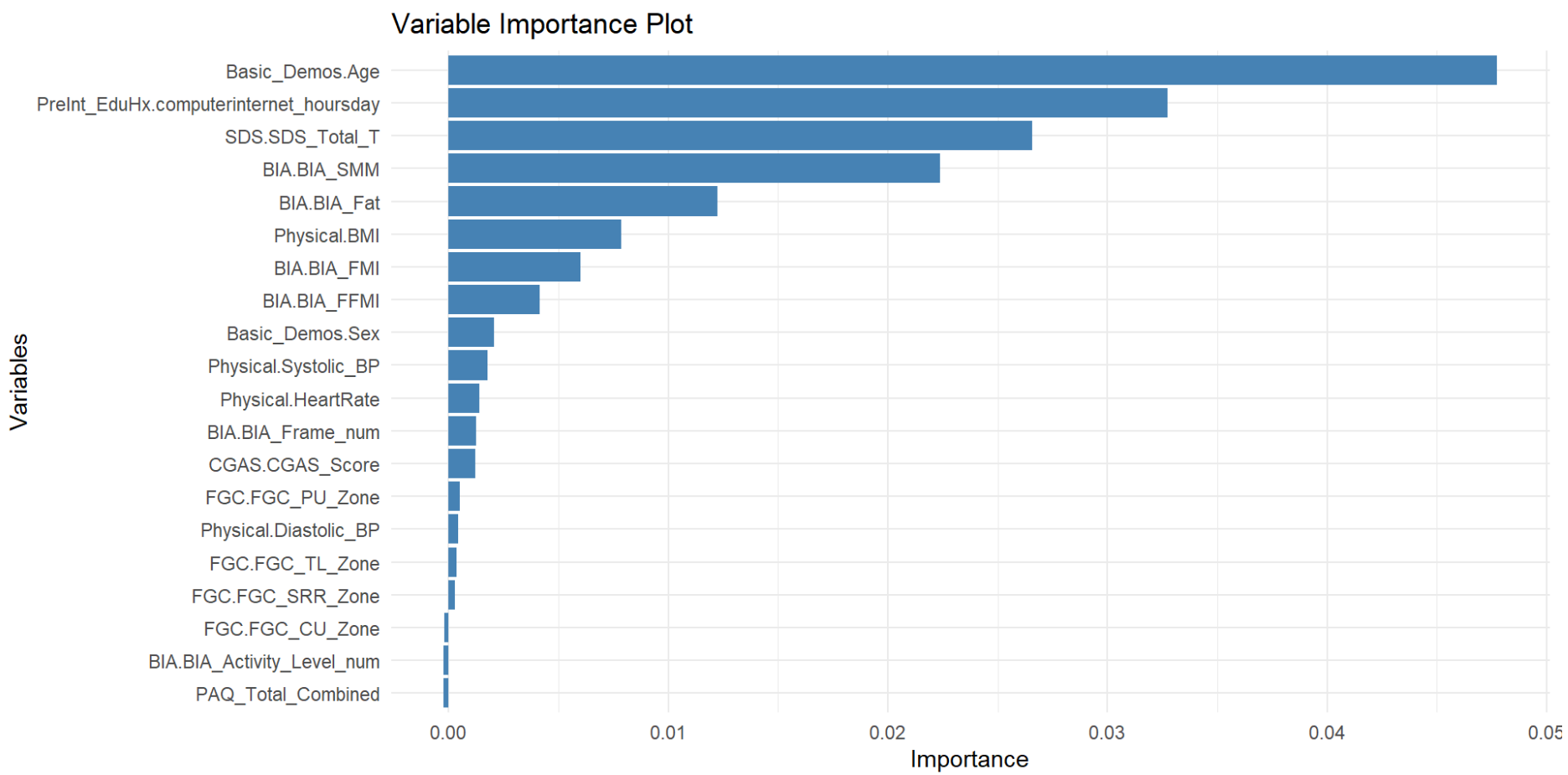
特徵重要性排名前三的變數：

- 每日使用電腦與網路平均時數 (PreInt_EduHx.computerinternet_hoursday)
- 性別(Basic_Demos.Sex)
- 活動水平(BIA_Activity_Level_num)

其中，性別和活動水平是負相關，表示男性的網路成癮程度可能高於女性，而活動水平越高，網路成癮程度傾向於降低

特徵重要性-Ordinal Forest

- Code



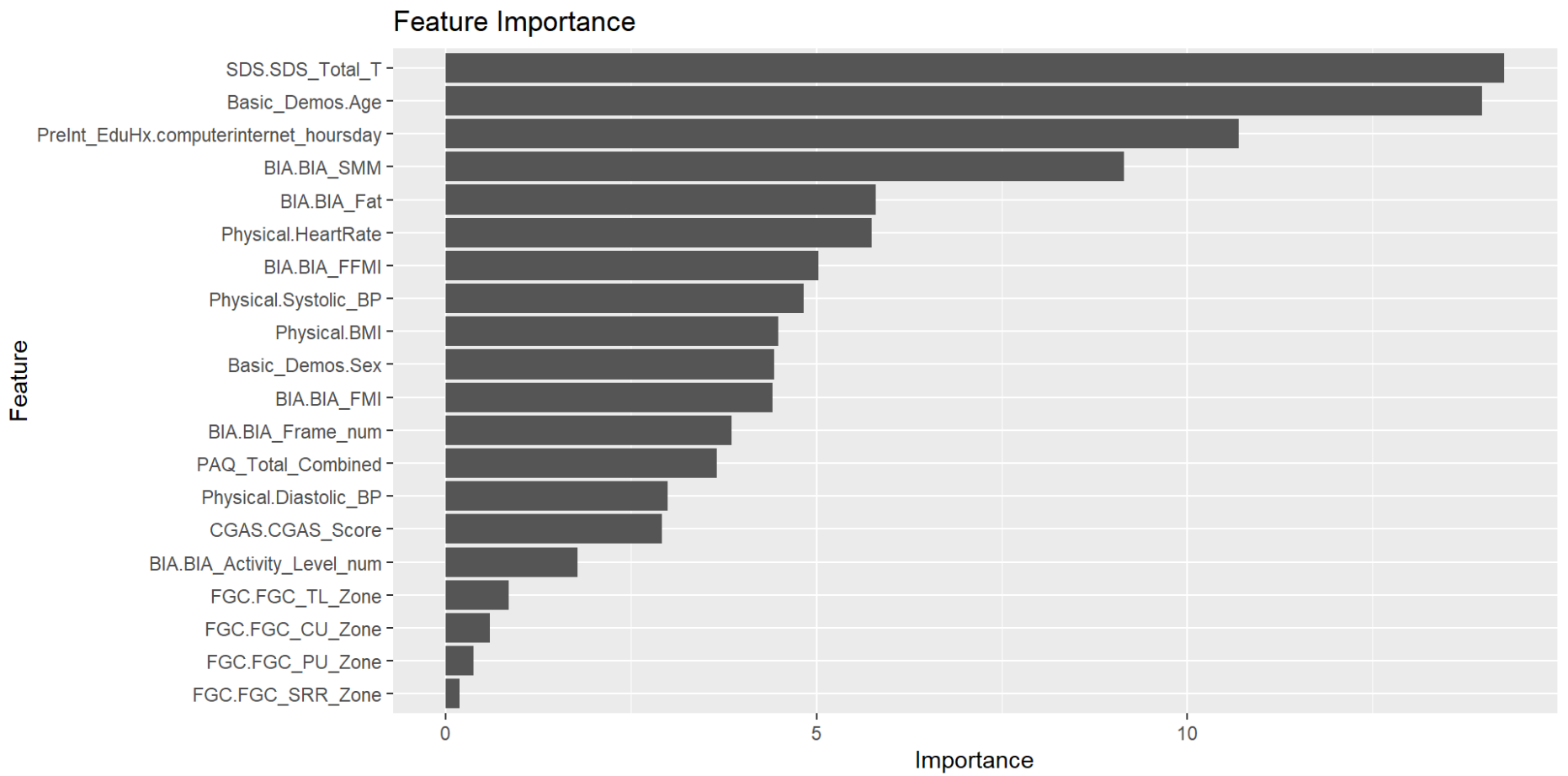
特徵重要性-Ordinal Forest

特徵重要性排名前三的變數：

- 年齡（Basic_Demos.Age）
- 每日使用電腦與網絡的平均時數（PreInt_EduHx.computerinternet_hoursday）
- 兒童睡眠障礙量表標準化總分（SDS.SDS_Total_T）

特徵重要性-CatBoost

- Code



特徵重要性-CatBoost

特徵重要性排名前三的變數：

- 兒童睡眠障礙量表標準化總分（SDS.SDS_Total_T）
- 年齡（Basic_Demos.Age）
- 每日使用電腦與網絡的平均時數
  （PreInt_EduHx.computerinternet_hoursday）

結論

綜合三個模型，可以推測

- 每日使用電腦與網絡的平均時數（PreInt_EduHx.computerinternet_hoursday）
- 年齡（Basic_Demos.Age）
- 兒童睡眠障礙量表標準化總分（SDS.SDS_Total_T）
- 性別 (basic_demos.sex)

對目標變數的影響可能較大，可以視為評估網路成癮的參考

工作分配

- Code

負責人	工作項目
廖芷萱	資料描述、資料前處理、口頭報告、書面報告製作
詹雅鈞	資料前處理、CatBoost、書面報告製作
李姿慧	資料前處理、Ordinal Forest、書面報告製作
謝沛恩	Ordinal Logistic Regression、書面報告製作
李敏榕	資料描述、簡報製作

參考資料

[1]Hancock, J. T., & Khoshgoftaar, T. M. (2020). CatBoost for big data: an interdisci plinary review. Journal of Big Data, 7(1), 94. https://doi.org/10.1186/s40537-020-00369 8

[2]J. K. Sayyad, K. Attarde and N. Saadouli(2024), “Optimizing e-Commerce Supply Chains With Categorical Boosting: A Predictive Modeling Framework,” in IEEE Access, vol. 12, pp. 134549-134567, 2024, doi: 10.1109/ACCESS.2024.3447756

[3]https://www.w3computing.com/articles/using-catboost-for-categorical-feature handling-in-machine-learning/

參考資料

[4]https://www.youtube.com/watch?v=KXOTSkPL2X4

[5]Hornung, R. (2017). Ordinal forests. Journal of Machine Learning Research, 18(159), 1–25.

[6]Institute for Digital Research and Education. (n.d.). Ordinal logistic regression in R. UCLA: Statistical Consulting Group. https://stats.oarc.ucla.edu/r/dae/ordinal-logistic regression/

參考資料

[7]Cheng Hua, Dr. Youn-Jeng Choi, Qingzhou Shi. (2021). Binary logistic regression. In Advanced regression techniques. Retrieved from Binary Logistic Regression (Bookdown)

[8]Shawn. (2024). 順序羅吉斯回歸 (Ordinal Logistic Regression)：介紹與解讀. Medium. Retrieved from Ordinal Logistic Regression 簡介與解讀

本組資料

https://github.com/H24101183/STATISTICAL-CONSULTING-FINAL-REPORT