子宮頸癌致病因素分析

310554049 王佳瑋
311554032 廖子濬

目錄

壹. 研究動機 **3**

貳. 資料來源 **3**

參. 資料敘述與分析 **4**

一. 資料敘述與預覽 4

a. Dataset Introduction and Data Distribution 4

b. Missing column and data 11

二. 資料關聯分析 12

共線性問題：Large F test but small T-test in some variables. 12

特徵篩選與羅吉斯回歸：Feature selection and Logistic regression 12

法一：對Full-model進行Reduce 14

法二：對已使用T-tset篩選過顯著之Variable進行Reduce 16

以機器模型對特定參數Dx.Cancer進行預測 19

模型一：XGBoostClassifer 20

模型二：CatBoostClassifer 22

三.特徵分群分析 24

特徵劃分：Feature Clustering 24

變數間的交互作用： Interaction between varivbles 25

肆. 討論與結論 **30**

壹. 研究動機

對於Association rule mining, 子宮頸癌是由HPV病毒感染所造成, 感染HPV目前無藥物可治療, 大部分是透過自身免疫力自行痊癒, 因此在免疫力低下的時候持續感染或發炎可能會導致發生子宮頸上皮細胞病變, 進一步變為子宮頸癌。有可能因為下述類似的原因, 如:不同型的HPV病毒可能會造成尖形濕疣(菜花)或是子宮頸癌;正確使用保險套可以阻擋HPV, HIV, 梅毒, B型肝炎的感染;MCV可以做為免疫力低下的指標, 等種種, 我們想知道說這些不同種的性病與子宮頸癌之間是否有關聯性。此外, 我們也希望透過資料集中提供的額外資料來了解當地的生活習慣及條件, 使我們可以更貼近當地的條件進行研究。

對於Classification, 由於在經濟條件較差的國家, 醫療資源並不充足, 每次的侵入式檢查都有非常高的風險, 且繁瑣和昂貴的切片檢查並不是每個人都負擔的起。因此, 我們希望藉由臨床特徵的資料對Biopsy 進行模型的訓練, 並利用這些模型來預測是否有做這些檢查的必要性, 因此我們目標是藉由分析臨床特徵來預測病人是否需要進行子宮頸切片, 並且期望在特異度能高於90%。

貳. 資料來源

我們所使用的資料集是由UCI所提供的Cervical cancer (Risk Factors) Data Set。這份dataset紀錄了由委瑞內拉某間醫院所統計罹患與疑似罹患子宮頸癌患者的病人屬性特徵與臨床特徵以及進行活體切片等子宮頸癌檢測後的結果。

網路上有很多與癌症相關的資料集, 但同樣是癌症, 這個dataset獨特在哪裡？ 子宮頸癌的切片檢查雖然程序較抹片複雜許多, 且成本較為昂貴, 但在台灣如抹片檢查有異狀者, 仍可以很普遍的接受切片檢查, 以決定是否進行前期治療, 大大降低了子宮頸癌的致死率。然而在經濟條件較差的國家, 如資料集蒐集的病例大多來自南美洲國家--委內瑞拉, 在當地並不是每個人都可以負擔的起如此繁瑣的切片檢查, 並且當地的醫療資源也不具有足夠的檢測能量對每個人進行檢測, 如果我們可以先做一個風險評估, 預測他們在做這些學理檢查的結果, 分析出他是高風險族群, 再對他進行檢查以確認是否罹癌, 把資源留給真正需要的人, 會比起因為後天限制而選擇不做下一步醫治, 可以救活更多的人。

參. 資料敘述與分析

一. 資料敘述與預覽

a. Dataset Introduction and Data Distribution

這份Dataset一共包含了858筆資料，每一筆資料所代表的是一名個別的病患，並且含有35種不同類型的特徵資料和1個target variable: Dx:Cancer，代表的是該筆資料患者是否罹癌，而其他欄位之資料也會在這個小節後半部加以解釋;資料所包含的feature和其資料類型如圖一所示:

Feature	Type	Feature	Type	Feature	Type
Age	int	STDs..number.	int	STDs.HPV	bool
Number.of.sexual.partners	int	STDs.condylomatosis	bool	STDs..Number.of.diagnosis	int
First.sexual.intercourse	int	STDs.cervical.condylomatosis	bool	STDs..Time.since.first.diagnosis	int
Num.of.pregnancies	int	STDs.vaginal.condylomatosis	bool	STDs..Time.since.last.diagnosis	int
Smokes	bool	STDs.vulvo.perineal.condylomatosis	bool	Dx.Cancer	bool
Smokes..years.	float	STDs.syphilis	bool	Dx.CIN	bool
Smokes..packs.year.	float	STDs.pelvic.inflammatory.disease	bool	Dx.HPV	bool
Hormonal.Contraceptives	bool	STDs.genital.herpes	bool	Dx	bool
Hormonal.Contraceptives..years.	float	STDs.molluscum.contagiosum	bool	Hinselmann	bool
IUD	bool	STDs.AIDS	bool	Schiller	bool
IUD..years.	float	STDs.HIV	bool	Citology	bool
STDs	bool	STDs.Hepatitis.B	bool	Biopsy	bool

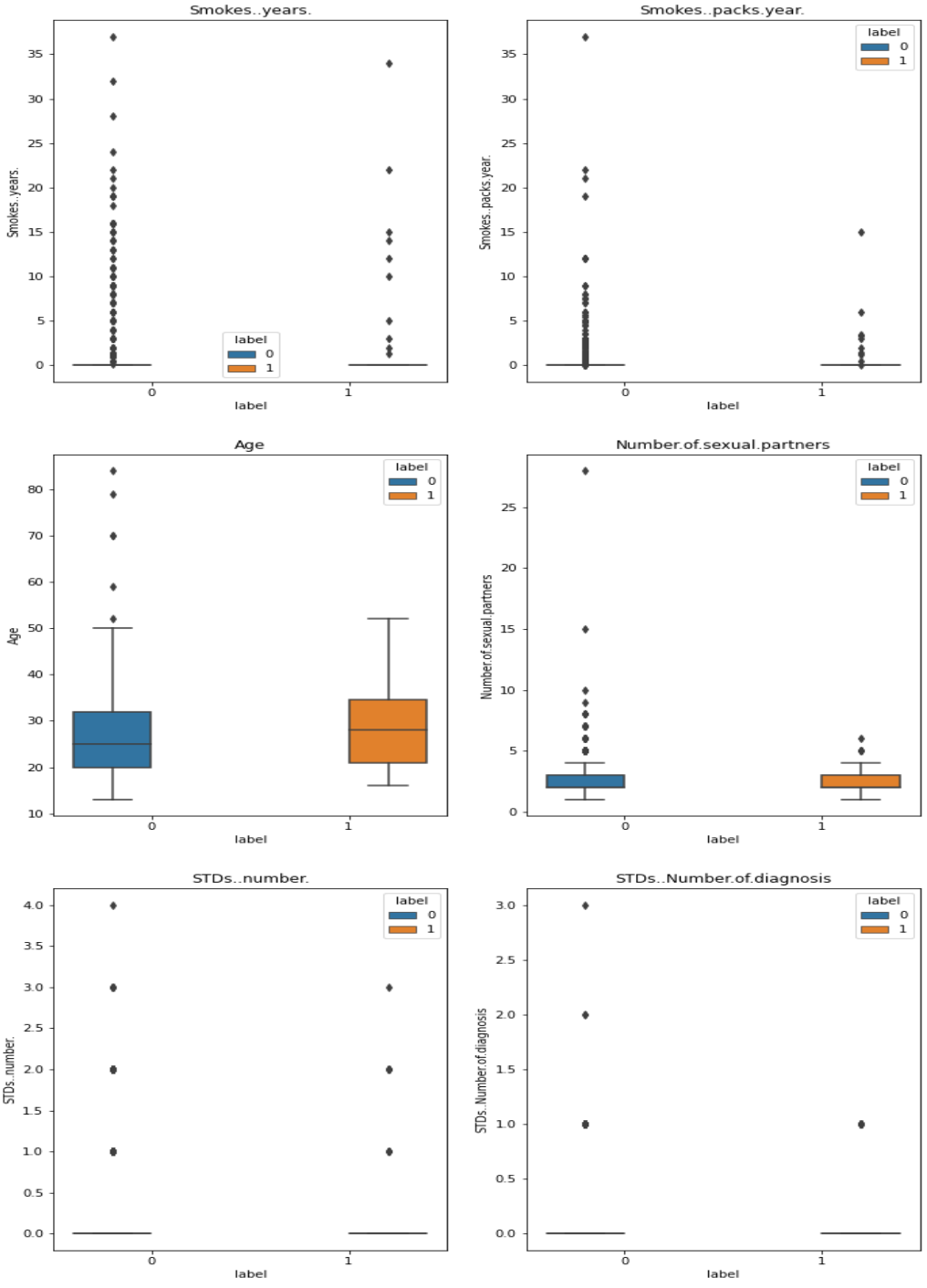
而其中，各個feature與其變數之解釋如下表所附:

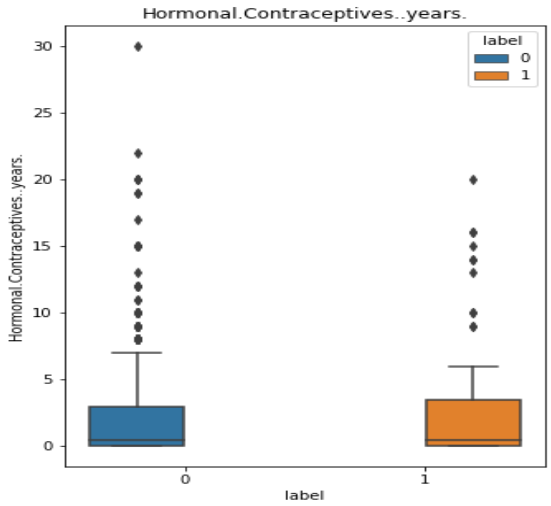
Age	年齡	STDs:syphilis	STDs:梅毒
Number of sexual partners	性伴侣的数量	STDs:pelvic inflammatory disease	STDs:盆腔炎症
First sexual intercourse (age)	第一次性交(年齡)	STDs:genital herpes	STDs:生殖器皰疹
Num of pregnancies	懷孕次數	STDs:molluscum contagiosum	性病:傳染性軟疣(Molluscum contagiosum

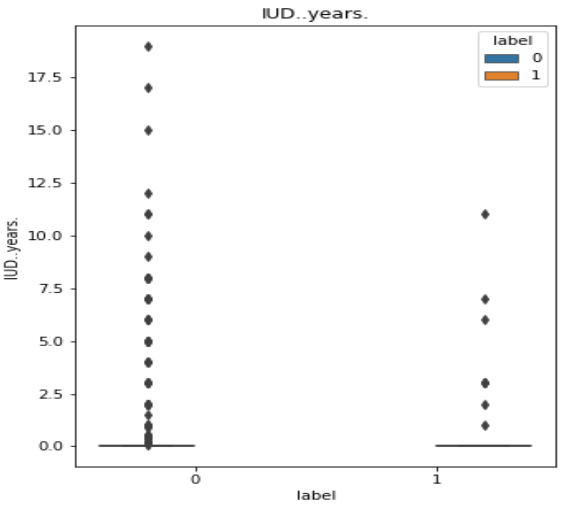
Smokes	吸煙	STDs:AIDS	STDs:AIDS
Smokes (years)	吸煙（年）	STDs:HIV	STDs:HIV
Smokes (packs/year)	吸煙（包/年）	STDs:Hepatitis B	STDs:B型肝炎
Hormonal Contraceptives	激素類避孕藥具	STDs:HPV	STDs:HPV
Hormonal Contraceptives (years)	激素避孕藥（年）	STDs: Number of diagnosis	性傳播疾病。診斷的數量
IUD	宮內節育器	Dx:Cancer	Dx:癌症
IUD (years)	宮內節育器（年數）	Dx:CIN	Dx:CIN 宮頸上皮內瘤樣病變（英語：Cervical intraepithelial neoplasia, CIN）
STDs	性傳播疾病	Dx:HPV	Dx:HPV
STDs (number)	STDs (number)	Dx	Dx
STDs:condylomatosis	STDs:尖銳濕疣	Hinselmann: target variable	Hinselmann: 目標變量
STDs:cervical condylomatosis	STDs:宮頸尖銳濕疣	Schiller: target variable	Schiller: 目標變量

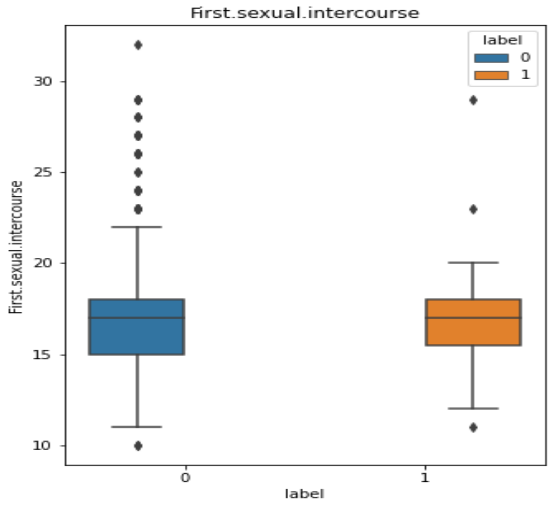
STDs:vaginal condylomatosis	STDs:陰道尖銳濕疣	Cytology: target variable	細胞學:目標變量
STDs:vulvo-perineal condylomatosis	STDs:會陰部尖銳濕疣	Biopsy: target variable	活組織檢查:目標變量

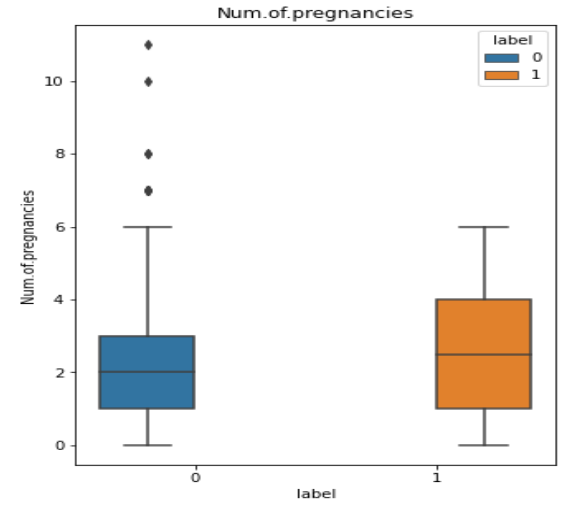
而各資料在這份資料集中具有以下的分佈, 以下是數值型類別之資料:

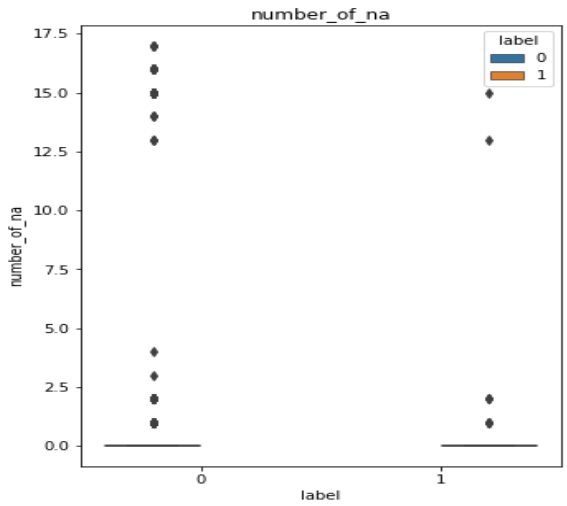



Hormonal.Contraceptives..years.
label
0
1
Hormonal.Contraceptives..years.
label

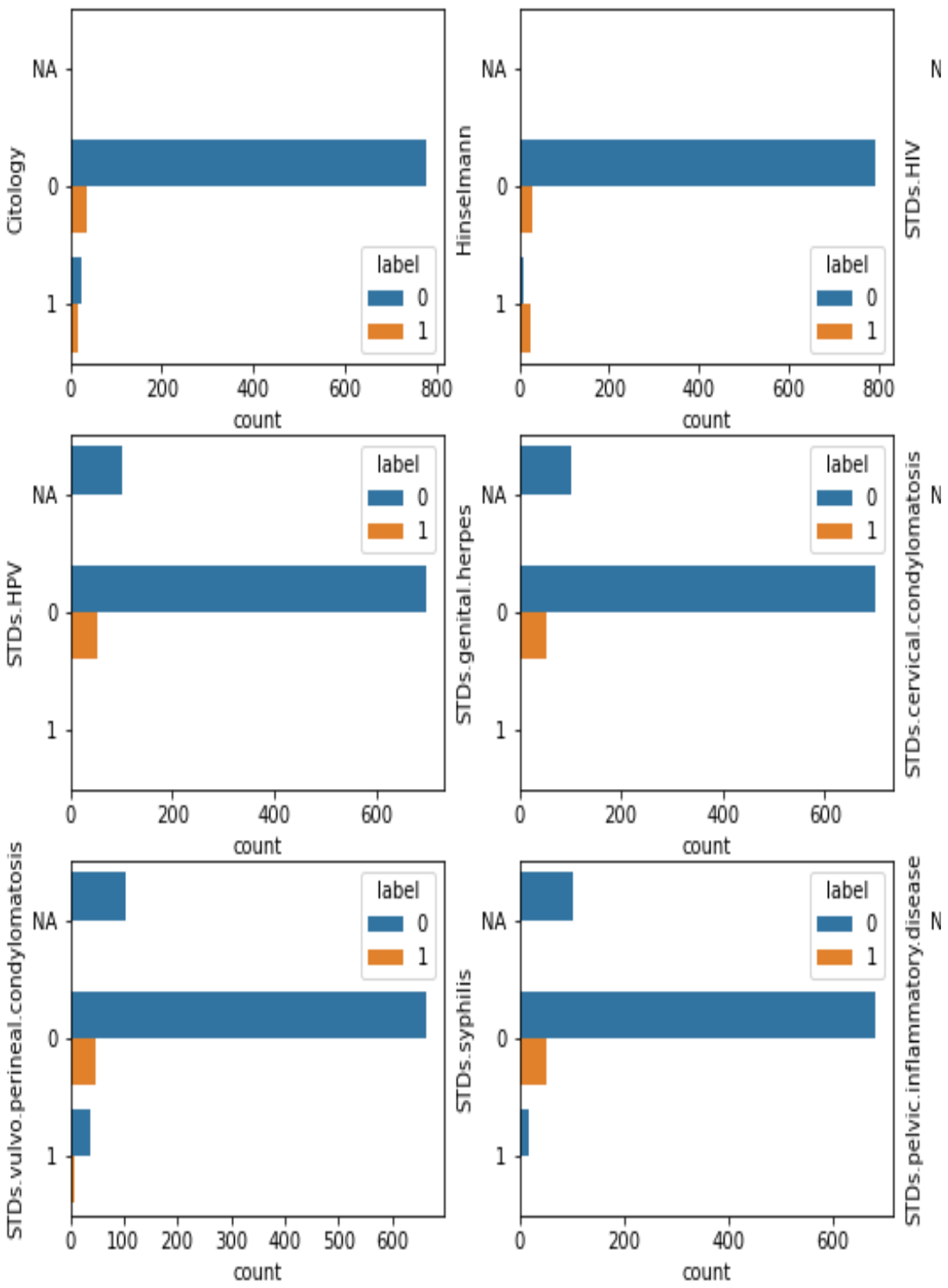

IUD..years.
label
0
1
IUD..years.
label

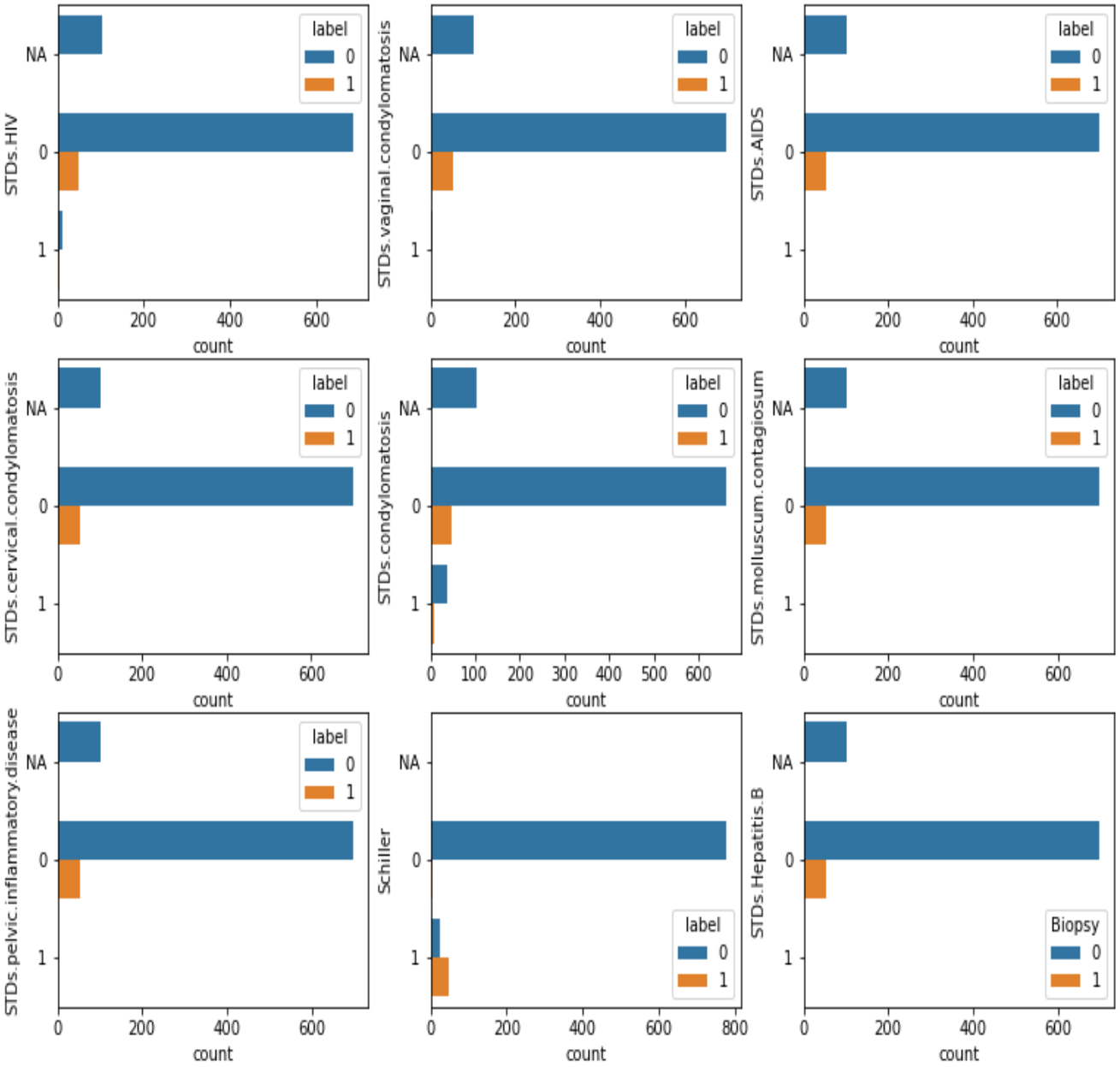

First.sexual.intercourse
label
0
1
First.sexual.intercourse
label


Num.of.pregnancies
label
0
1
Num.of.pregnancies
label


number_of_na
label
0
1
number_of_na
label

接下來是類別特徵, 從圖中可以看到有些特定特徵在整個資料集中都是單一值, 像是STDs:AIDS欄位中只有0的值。以下是布林型別之資料分佈:





在上述圖中可以看到，在數值類型的特徵中，有許多特徵包含outliers，但因我們所使用的資料集本身資料點不多，所以在訓練模型的過程中並未移除outliers。事實上我們在做了outlier detection去檢查boundary cases後，發現有265筆data被定義為outliers，這對我們總共858筆資料來說，實則約1/3的資料量，很明顯是母體數的不足所造成的，或許在fit模型的階段中，我們認為不應該輕易的將其當作outliers，而需將其列入考慮。

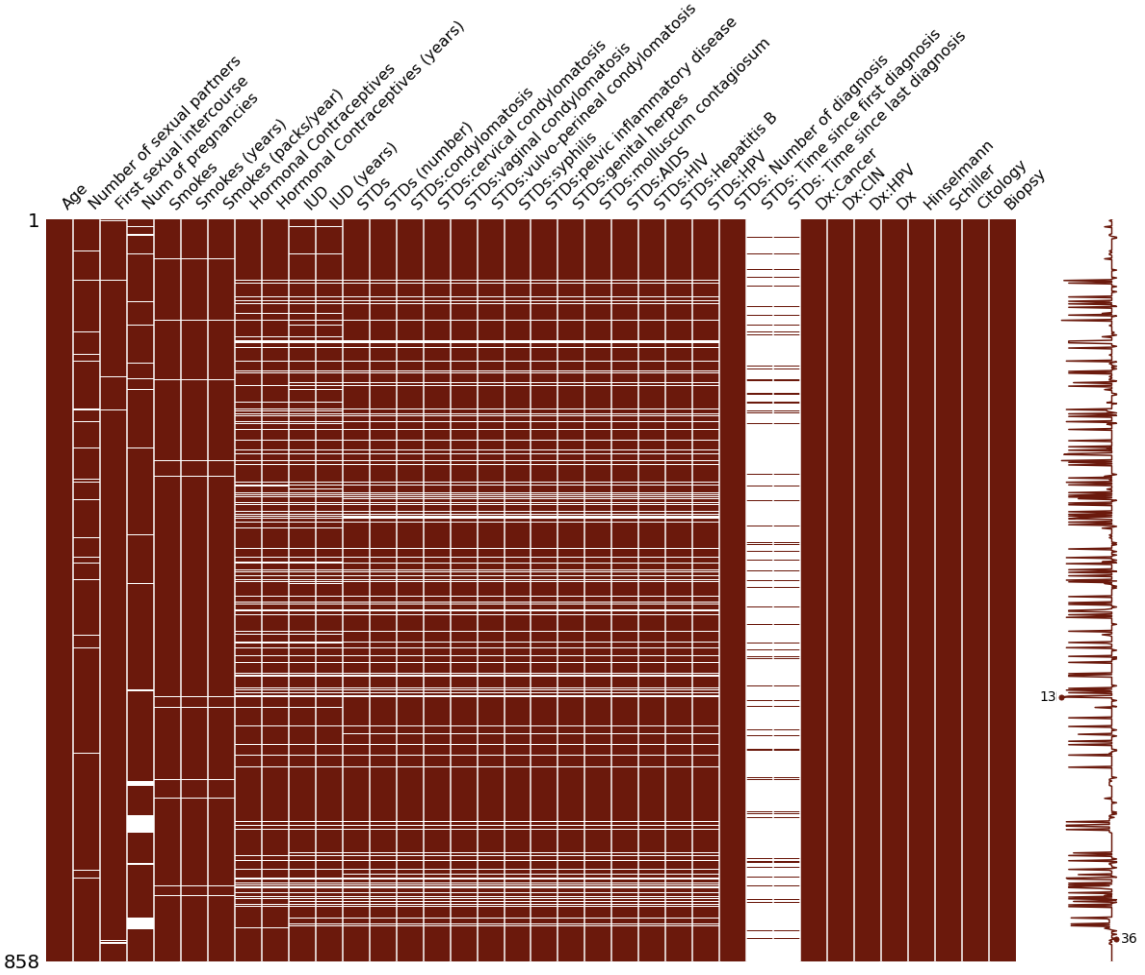
而關乎資料數值, 各欄位在四個百分位距的分佈如下表所示:

Attribute	Min	1st Qu	Median	Mean	3rd Qu	Max	NA's
Age	13	20	25	26.82	32	84	0
Number.of.sexual.partners	1.000	2.000	2.000	2.528	3.000	28.000	26
First.sexual.intercourse	10	15	17	17	18	32	7
Num.of.pregnancies	0.000	1.000	2.000	2.276	3.000	11.000	56
Smokes	0.0000	0.0000	0.0000	0.1456	0.0000	1.0000	13
Smokes..years.	0.00	0.00	0.00	1.22	0.00	37.00	13
Smokes..packs.year.	0.0000	0.0000	0.0000	0.4531	0.0000	37.0000	13
Hormonal.Contraceptives	0.0000	0.0000	1.0000	0.6413	1.0000	1.0000	108
Hormonal.Contraceptives..years.	0.000	0.000	0.500	2.256	3.000	30.000	108
IUD	0.000	0.000	0.000	0.112	0.000	1.000	117
IUD..years.	0.0000	0.0000	0.0000	0.5148	0.0000	19.0000	117
STDs	0.0000	0.0000	0.0000	0.1049	0.0000	1.0000	105
STDs..number.	0.0000	0.0000	0.0000	0.1766	0.0000	4.0000	105
STDs.condylomatosis	0.00000	0.00000	0.00000	0.05843	0.00000	1.00000	105
STDs.cervical.condylomatosis	0	0	0	0	0	0	105
STDs.vaginal.condylomatosis	0.00000	0.00000	0.00000	0.00531	0.00000	1.00000	105
STDs.vulvo.perineal.condylomatosis	0.00000	0.00000	0.00000	0.0571	0.00000	1.00000	105
STDs.syphilis	0.0000	0.0000	0.0000	0.0239	0.0000	1.0000	105
STDs.pelvic.inflammatory.disease	0.00000	0.00000	0.00000	0.00133	0.00000	1.00000	105
STDs.genital.herpes	0.00000	0.00000	0.00000	0.00133	0.00000	1.00000	105
STDs.molluscum.contagiosum	0.00000	0.00000	0.00000	0.00133	0.00000	1.00000	105
STDs.AIDS	0	0.00000	0.00000	0.00000	0	0.00000	105
STDs.HIV	0.0000	0.0000	0.0000	0.0239	0.0000	1.0000	105
STDs.Hepatitis.B	0.00000	0.00000	0.00000	0.00133	0.00000	1.00000	105
STDs.HPV	0.00000	0.00000	0.00000	0.00266	0.00000	1.00000	105
STDs..Number.of.diagnosis	0.00000	0.00000	0.00000	0.08741	0.00000	3.00000	787
STDs..Time.since.first.diagnosis	1.000	2.000	4.000	6.141	8.000	22.000	787
STDs..Time.since.last.diagnosis	1.000	2.000	3.000	5.817	7.500	22.000	787
Dx.Cancer	0.00000	0.00000	0.00000	0.02098	0.00000	1.00000	0
Dx.CIN	0.00000	0.00000	0.00000	0.01049	0.00000	1.00000	0
Dx.HPV	0.00000	0.00000	0.00000	0.02098	0.00000	1.00000	0

Attribute	Min	1st Qu	Median	Mean	3rd Qu	Max	NA's
Dx	0.00000	0.00000	0.00000	0.02797	0.00000	1.00000	0
Hinselmann	0.00000	0.00000	0.00000	0.04079	0.00000	1.00000	0
Schiller	0.00000	0.00000	0.00000	0.08625	0.00000	1.00000	0
Citology	0.00000	0.00000	0.00000	0.05128	0.00000	1.00000	0
Biopsy	0.00000	0.00000	0.00000	0.06410	0.00000	1.00000	0

b. Missing column and data

在這份資料集中不乏存在缺失的欄位，我們在之後的統計中勢必要對其進行處理。而在下圖中我們可以觀察NA欄位的分佈，多數在STDs..Time.since.first.diagnosis, STDs..Time.since.last.diagnosis這兩個欄位，在日後分析中比起填入值，我們更偏向將此二欄位進行移除。而其餘缺失欄位會再根據相應的分析做調整。



二. 資料關聯分析

共線性問題：Large F test but small T-test in some variables.

首先，在我們的測試結果中，遇到了F test 很大但 T-test 很小的情況，也就是共線性的問題，自我們的資料集中包含了許多的性病，有各種性病的各自的統計資料還有每個患者性病總數，這幾項資料就會有共線性的問題。因此，我們會用不同的方式來處理這些資料，一種是我們會將統計結果的資料移除如 STDs.Numbers 來看是哪種性病會導致子宮頸癌，接著再將統計的資料放回去並移除個別的性病資料做另一次模型，看看是否性病的數量才是影響結果比較重要的原因而非疾病的差異。

在T-Test中顯著的類別分別是：Age, First.sexual.intercourse, STDs.condylomatosi, STDs.cervical.condylomatosis, STDs.vaginal.condylomatosis, STDs.vulvo.perineal.condylomatosis, STDs.syphilis, STDs.AIDS, STDs.HIV, Dx.CIN, Dx.HPV, Dx, Schiller, Biopsy，但這樣仍有許多的變數，我們想再下一階段建立足夠簡單的模型來預測我們感興趣的變數Dx.Cancer，這樣的變數數量對我們來說仍然是有點多，預期能夠用到少的解釋變數。

特徵篩選與羅吉斯回歸：Feature selection and Logistic regression

在做完T-Test中我們得到了相對重要影響Dx.Cancer的變數，但我們也想過，若是直接從Full model直接做backward mwthod去extrace important variables是否會得到同樣的結果？其表現會比從做完T-Test之參數再去篩選下來，表現更好或是更差？由於這個模型原先有35個變數，我們盡可能地希望可以得到一個簡單的model，但又不希望在分析的部分過於複雜，於是我們設計了兩種方式：

法一：對Full model直接使用StepAIC之套件，選取重要的變數。這邊使用的挑選方式是Backward method，逐次刪減直到找到最小的AIC。

法二：對在經過T-Test之後之重要變數再加以篩選，我們用到的方法與法一相似，用StepAIC之套件採Backward method挑選。

而在這邊為了不影響我們在最後在模型評估遇到預測資料與原先資料集levels不同的問題，我們必須先對原有資料的NA值做處理。以下是我們的處理方法：

```
Unset
# Remove the specified features from the dataset
cervical_cancer <- cervical_cancer[, !(names(cervical_cancer) %in%
c("STDs..Time.since.first.diagnosis", "STDs..Time.since.last.diagnosis", "STDs.AIDS",
"STDs.cervical.condylomatosis"))]

# Remove rows with missing values from the dataset
cervical_cancer_clean <- na.omit(cervical_cancer)
```

根據前面的資料預覽，將無用（STDs.AIDS, STDs.cervical.condylomatosis）/ 空缺資料過多（STDs..Time.since.first.diagnosis, STDs..Time.since.last .diagnosis）的Feature先行移除，再將有缺失值的單筆資料移出資料集；最後我們得到了從原先 858*36 的資料集變為 668*32大小的資料集，其中無NA欄位。

接下來我們切分資料以獲取我們所需的訓練集以及測試集：

```
Unset
library(caTools)
set.seed(123)
split <- sample.split(cervical_cancer_clean$Dx.Cancer, SplitRatio = 0.7)
train <- cervical_cancer_clean[split, ]
test <- cervical_cancer_clean[!split, ]
```

```
Unset
summary(test)
Dx.Cancer
0:195
1:5
```

我們將訓練集以及測試集沿著7:3的比例去做分割，我們可以觀察新得到之變數 train, test。接下來我們將展示兩種方法所得到之結果。

法一：對Full-model進行Reduce

在最終的變數篩選我們得到的變數有Smokes + Smokes..years. + Dx.CIN + Dx.HPV + Dx，我們觀察從這些變數建立模型得到的結果：

```
Unset
Call:
glm(formula = Dx.Cancer ~ Smokes + Smokes..years. + Dx.CIN +
    Dx.HPV + Dx, family = binomial, data = cervical_cancer_clean)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)       -26.741  16510.624  -0.002    0.999
Smokes           -132.009  64003.091  -0.002    0.998
Smokes..years.      4.063   1739.087   0.002    0.998
Dx.CIN            -50.701 218751.419   0.000    1.000
Dx.HPV            133.707  51454.490   0.003    0.998
Dx                 50.876  76485.399   0.001    0.999

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1.5838e+02  on 667  degrees of freedom
Residual deviance: 1.5521e-08  on 662  degrees of freedom
AIC: 12

Number of Fisher Scoring iterations: 25
```

從這份報表中我們可以看到，在p value的部分是較為差強人意，且在Std. Error有著很高的殘差。在Warning中提到了有關模型無法收斂的問題，這或許是因為與資料分佈的關係有所關聯，在Dx.Cancer，我們所預測的變數中，這項變數為1的頻率較為稀少，是因為貼近現實情況，收集該地醫院與癌症相關之病例，但實際確診病例少，是一種imbalanced data.

接著我們觀察這份資料在模型上的表現，我們主要觀察的對象是使用Confusion Matrix中所得出的型一錯誤與型二錯誤。

```
Unset
Warning: glm.fit: algorithm did not convergeWarning: glm.fit: fitted probabilities numerically 0 or 1 occurredWarning: prediction from rank-deficient fit; attr(*, "non-estim") has doubtful casesConfusion Matrix and Statistics

          Reference
Prediction   0   1
         0 194   0
         1   1   5

               Accuracy : 0.995
                 95% CI : (0.9725, 0.9999)
    No Information Rate : 0.975
    P-Value [Acc > NIR] : 0.03875

                  Kappa : 0.9065

 Mcnemar's Test P-Value : 1.00000

            Sensitivity : 0.9949
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.8333
             Prevalence : 0.9750
         Detection Rate : 0.9700
   Detection Prevalence : 0.9700
      Balanced Accuracy : 0.9974

       'Positive' Class : 0
```

在這份報表中我們可以看出，準確度，信賴區間，敏感度和特異度。對於準確度與信賴區間，報表中顯示了模型的準確度和95%的信賴區間。準確度表示模型正確預測的比例，這裡的準確度為0.995。信賴區間則是對準確度的估計進行統計上的區間估計，這裡的信賴區間為(0.9725, 0.9999)，表示我們對準確度有95%的信心介於這個區間內。敏感度和特異度：報表中顯示了模型的敏感度（Sensitivity）和特異度（Specificity）。敏感度表示模型正確檢測出正例的能力，這裡的敏感度為0.9949，表示模型能夠很好地檢測出類別1的樣本。

法二:對已使用T-tset篩選過顯著之Variable進行Reduce

回顧在T-Test中顯著的類別分別是:Age, First.sexual.intercourse, STDs.condylomatosi, STDs.cervical.condylomatosis, STDs.vaginal.condylomatosis, STDs.vulvo.perineal.condylomatosis, STDs.syphilis, STDs.AIDS, STDs.HIV, Dx.CIN, Dx.HPV, Dx, Schiller, Biopsy。但我們為求簡單模型, 這個數量的變數讓我們仍有挑選的空間。我們再次使用StepAIC對這個已經篩選過的候選features再度精簡, 而得到以下的結果, 我們先展示一組Attr = Age, 是一具有顯著結果的T-test, 他有這樣的格式:

```
Unset
Age t-test result:

        Welch Two Sample t-test

data:  with_cancer[[col]] and without_cancer[[col]]
t = 3.331, df = 17.776, p-value = 0.003768
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
  2.411016 10.666762
sample estimates:
mean of x mean of y
 33.22222  26.68333
```

而我們將所有具有顯著結果的T-test整理成表格, 如下圖所示:

Attribute	t	df	p-value	confidence interval upper	confidence interval lower	mean of x	mean of y
Age	3.331	17.776	0.003768	10.666762	2.411016	33.22222	26.68333
First.sexual.intercourse	2.8533	18.642	0.0103	2.272522	0.347859	18.27778	16.96759
STDs.condylomatosis	-6.8365	734	1.708e-11	-0.04267320	-0.07705469	0.00000000	0.05986395
STDs.vaginal.condylomatosis	-2.0041	734	0.04543	-0.0001110609	-0.0107732928	0.000000000	0.005442177
STDs.vulvo.perineal.condylomatosis	-6.7535	734	2.936e-11	-0.04149683	-0.07550997	0.0000000	0.0585034
STDs.syphilis	-4.2926	734	2.003e-05	-0.01328961	-0.03568998	0.0000000	0.0244898
STDs.HIV	-4.2926	734	2.003e-05	-0.01328961	-0.03568998	0.0000000	0.0244898
Dx.CIN	-3.0144	839	0.002652	-0.003737803	-0.017690769	0.00000000	0.01071429
Dx.HPV	11.628	17.017	1.609e-09	1.0473487	0.7256672	0.888888889	0.002380952
Dx	7.5903	17.047	7.281e-07	0.9787113	0.5530347	0.77777778	0.01190476
Schiller	2.6064	17.213	0.01831	0.5591256	0.0591284	0.3888889	0.0797619
Biopsy	2.3993	17.171	0.02804	0.51664058	0.03335942	0.33333333	0.05833333

對這些變數再做Feature Seection Backward method後之結果:

```
Unset
Call:  glm(formula = Dx.Cancer ~ First.sexual.intercourse + Dx.HPV +
    Dx, family = binomial, data = train)

Coefficients:
          (Intercept)  First.sexual.intercourse                Dx.HPV
             -314.848                     9.234                171.661
                   Dx
              157.866

Degrees of Freedom: 333 Total (i.e. Null);  330 Residual
Null Deviance:              75.51
Residual Deviance: 4.639e-08      AIC: 8
```

可以觀察到剩下First.sexual.intercourse, Dx.HPV與Dx作為模型的評斷標種。至於最後得到的模型, 我們觀察下列報表得知, 這樣的模型在顯著性上並沒有特別突出, 但我們仍然使用Confusion Matrix 進行評估:

```
Unset
Call:
glm(formula = stepAIC_formula, family = binomial, data = train)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                       -212.119  71592.464  -0.003    0.998
Age                                1.525    646.407   0.002    0.998
First.sexual.intercourse           4.422    1764.867   0.003    0.998
STDs.condylomatosis                110.279 409070.792   0.000    1.000
STDs.vaginal.condylomatosis        -26.142 196670.006   0.000    1.000
STDs.vulvo.perineal.condylomatosis -75.894 397728.904   0.000    1.000
STDs.syphilis                      1.169 625130.906   0.000    1.000
STDs.HIV                           21.805 594243.834   0.000    1.000
Dx.CIN                             -23.838 371251.387   0.000    1.000
Dx.HPV                             103.154  33541.428   0.003    0.998
Dx                                 104.925 107457.733   0.001    0.999
Schiller                           38.999 154347.508   0.000    1.000
Biopsy                             -28.768 168484.142   0.000    1.000

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 7.5514e+01  on 333  degrees of freedom
Residual deviance: 1.9795e-08  on 321  degrees of freedom
AIC: 26

Number of Fisher Scoring iterations: 25
```

```
Unset


Confusion Matrix and Statistics


          Reference
Prediction   0   1
         0 193   0
         1   2   5


               Accuracy : 0.99
```

```
               95% CI : (0.9643, 0.9988)
    No Information Rate : 0.975
    P-Value [Acc > NIR] : 0.1215

                  Kappa : 0.8283

 Mcnemar's Test P-Value : 0.4795

            Sensitivity : 0.9897
            Specificity : 1.0000
         Pos Pred Value : 1.0000
         Neg Pred Value : 0.7143
             Prevalence : 0.9750
         Detection Rate : 0.9650
   Detection Prevalence : 0.9650
      Balanced Accuracy : 0.9949

       'Positive' Class : 0
```

在這份報表中，準確率的95%信賴區間為(0.9785, 0.9993)，而相對於使用Full-model進行篩選，這份資料在型一錯誤與型二錯誤的表現上是相對較好，但相同的，模型顯著性也不高，這使我們開始思考，從一些機器學習的角度進行切入。

以機器模型對特定參數Dx.Cancer進行預測

模型一：XGBoostClassifer

XGBoost是一種梯度提升框架，用於解決分類和回歸問題。它是由陳天奇在2014年開發的，並且在機器學習競賽中取得了顯著的成功。XGBoost是「Extreme Gradient Boosting」的縮寫。

XGBoost通過集成多個弱學習器（通常是決策樹）來構建一個強大的預測模型。它採用了梯度提升算法，每次迭代都會優化損失函數，以逐步減小預測誤差。與傳統的梯度提升算法相比，XGBoost引入了一些創新和優化技術，使得它在效率和準確性上都表現出色。

以下是XGBoost的一些主要特點和優勢：

1. 正歸化：XGBoost通過正歸化技術（如L1和L2正歸化）來控制模型的複雜度，防止過擬合。
2. 自動處理缺失值：XGBoost能夠自動處理缺失值，無需對缺失值進行額外的處理。
3. 內置交叉驗證：XGBoost內置了交叉驗證功能，可以幫助選擇最佳的模型參數，提高模型的泛化能力。
4. 特徵重要性評估：XGBoost可以計算特徵的重要性，幫助理解模型的預測過程和特徵的貢獻程度。
5. 並行處理：XGBoost使用了並行計算技術，能夠快速處理大規模數據集。
6. 可擴展性：XGBoost支持分布式計算，可以在集群上進行訓練和預測，適用於處理大規模數據和高維特徵的場景。

XGBoost是一種功能強大且高效的機器學習算法，在各種數據科學任務中廣泛應用，包括預測建模、排名、回歸和異常檢測等領域。對於我們的資料，我們希望通過XGBoost應用決策樹的特性，得到一個在模型解釋上較有根據的結果。

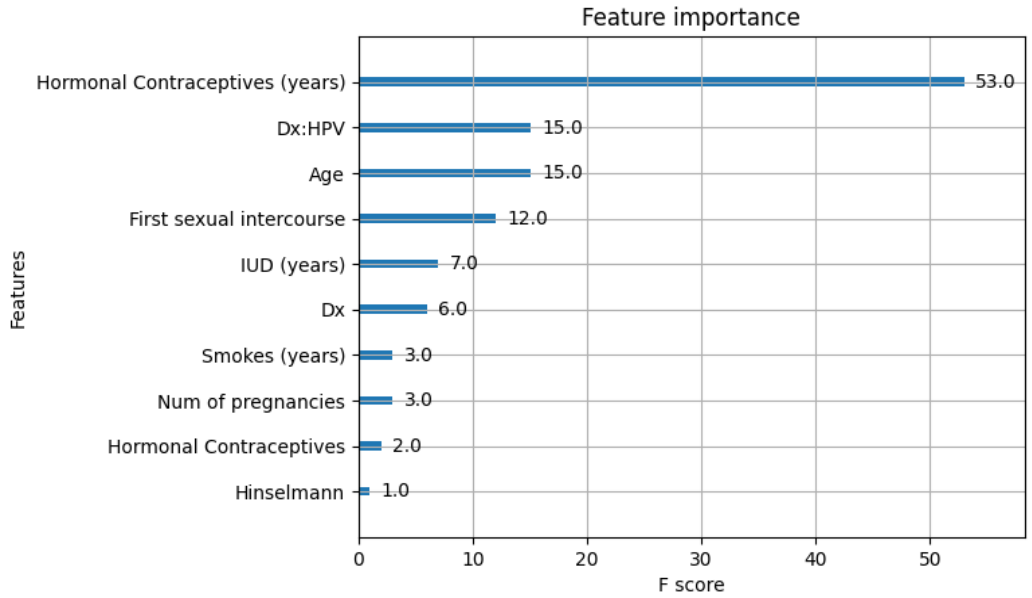
```
Unset
Train Result:

Accuracy Score: 99.79%
Precision Score: 100.00%
Recall Score: 90.00%
F1 score: 97.31%
Confusion Matrix:
 [[457   0]
 [  1   9]]

Test Result:
Accuracy Score: 99.00%
Precision Score: 100.00%
Recall Score: 71.43%
F1 score: 91.41%
Confusion Matrix:
 [[194   0]
 [  2   5]]
```

在這邊測試訓練的模型，我們分別將訓練集與測試集丟入模型中測試模型表現。在訓練集放回模型中的表現中，我們可以看到預測結果並不是完全的100% Accuracy。或許會覺得奇怪，為何用來訓練的資料照理來說是已經看過的資料，還會發生預測錯誤的情況？那是因為模型具有防止過擬合的特性，同時力求簡單的model，所必要作出的犧牲。但當然這點在CatBoost會有所改進，我們可以從待會的分析數據中觀測到。

對於預測集，我們沿用了在RStudio套用Logistic Regression的特性，將訓練集與資料集切成7:3的比例。而在測試集預測的表現中，在Accuracy會有好表現是可以預期的；在這樣的imbalanced data下，模型可以輕易地預測出較為多數的結果，但是否在較少出現的case（如：罹癌的人）預測成功，這點才是我們該去關注的。Recall是一個很好的指標，即事實為真的樣本中有幾個是預測正確的。在這個模型中有71.43%的Recall，代表我們可以找出大部分罹癌的樣本，但或許還不夠好。ㄧ樣，在待會的CatBoost會在這邊上有所改善。



從模型的分析組成，我們可以看到主宰決定Dx:Cancer的Attribute為Hormonal Contrapositive，這與我們在Full-model Selection有Smokes + Smokes..years. + Dx.CIN + Dx.HPV + Dx，在T-test reduction得到First.sexual.intercourse，Dx.HPV與Dx有點不同，可能共同點是：這些模型都覺得Dx:HPV是一個重要的變數，而之間在決定重要變數仍有一些差異，基於F-Score或是基於樹狀模型所定，有異曲同工之妙。

模型二：CatBoostClassifer

Catboost是一種基於梯度提升算法的機器學習框架，用於解決分類和回歸問題。它由Yandex團隊開發，並在2017年開源發佈。Catboost的名稱中的"Cat"代表"Category"，因為它在處理分類變量（離散特徵）方面表現出色。

Catboost和XGBoost都是強大的梯度提升框架，它們在許多方面都表現出色。以下是Catboost相對於XGBoost的一些優勢：

1. 處理分類變量：Catboost在處理分類變量方面具有內建的優勢。它能夠自動處理類別特徵的編碼和缺失值，並在模型訓練過程中有效地利用這些特徵。相比之下，XGBoost需要手動對分類特徵進行編碼和處理。
2. 自動特徵縮放：Catboost能夠自動進行特徵縮放，減少特徵值之間的偏差對模型訓練的影響。這樣可以簡化數據預處理的步驟，並提高模型的訓練效率。而XGBoost在使用前需要手動對特徵進行縮放。
3. 對稀疏數據的優化：Catboost對稀疏數據集進行了優化，能夠高效地處理這類數據。它使用了特定的數據結構和算法，減少了內存的使用和計算的複雜性。相比之下，XGBoost對稀疏數據的處理相對較慢。
4. 內置交叉驗證：Catboost內置了交叉驗證功能，可以幫助選擇最佳的超參數，並評估模型的性能。這樣可以簡化模型調優的過程，減少了額外的編碼和計算。
5. 支持GPU加速：Catboost支持在GPU上進行訓練和預測，可以顯著加快模型的訓練速度。XGBoost也支持GPU加速，但Catboost在GPU上的性能通常更好。

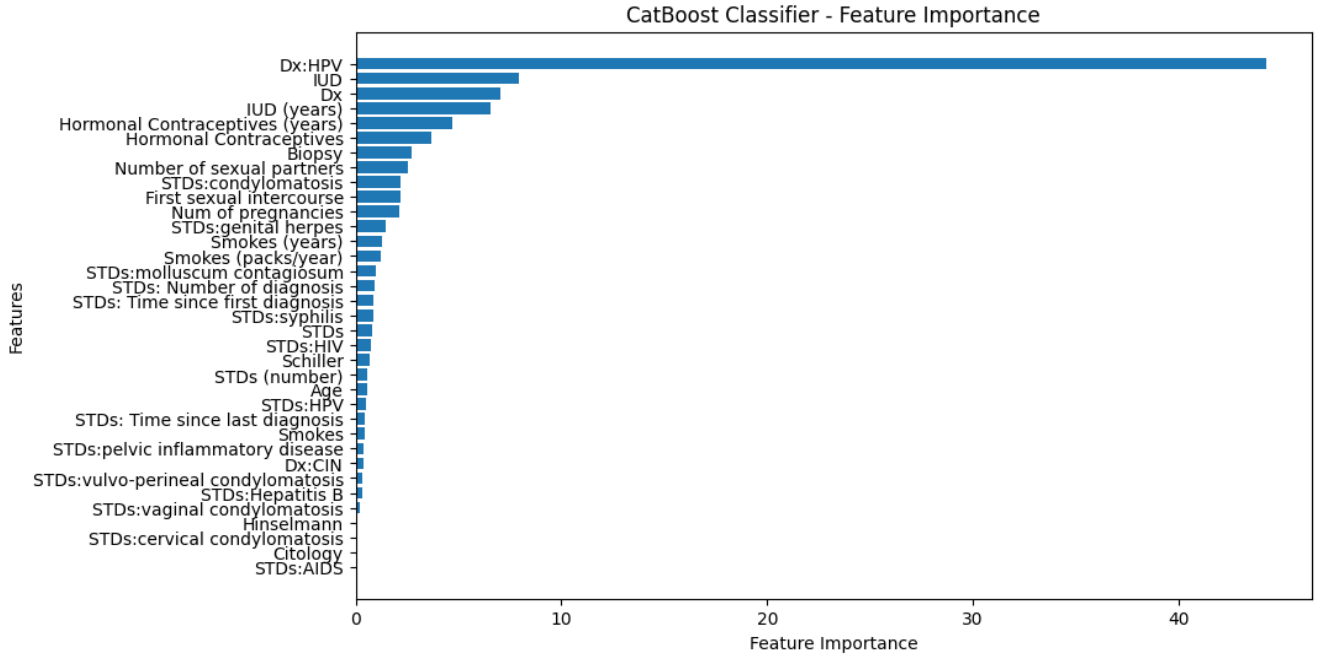
需要注意的是，Catboost和XGBoost在不同情況下的性能可能有所差異，具體取決於數據集的特點和任務的要求。因此，在選擇模型時，最好根據具體情況進行比較和評估，選擇最適合的框架。對於我們的訓練來說，因為這份資料有大量的NA（空缺欄位），且資料量不大，若是將這些資料移除，能判斷/影響模型組成的資料或許就會因此缺失，所以CatBoost可以在這些資料有部分缺失的情況下，仍用這些有限的資料進行判讀，我認為這是一個優點，也是我挑選這模型的原因。

```
Unset
Train Result:
Accuracy Score: 100.00%
Precision Score: 100.00%
Recall Score: 100.00%
F1 score: 100.00%
Confusion Matrix:
 [[585   0]
 [  0  15]]

Test Result:
```

```
Accuracy Score: 99.61%
Precision Score: 75.00%
Recall Score: 100.00%
F1 score: 92.76%
Confusion Matrix:
 [[254   1]
 [  0   3]]
```

首先，我們先觀察在Confusion Matrix上的表現，我們可以看到在訓練集去擬和訓練出來的模型有較好的擬合度，但我們同時又會擔心在這份數據上會有過擬合的情形，在樹模型中較多分支，造成訓練出的模型過於複雜。但隨後觀察測試集得出的預測結果，我們發現在Accuracy/Precision中的表現都比XGBoost好，推測他的確有一個比比XGBoost較為複雜的模型，但主因在於能加以運用於判別的資料量變多了，在此同時表現也變好了，同時也符合了我們的假設，CatBoost這樣的模型適合此性質資料集的訓練與預測。



在這份模型中認為Dx:HPV是一項重要的變數，與前三個模型所提出的分析結果相似，他們都認為Dx:HPV是一個對於Dx:Cancer重要的變數，但令人意外的是，他在前幾個重要變數與XGBoost相同，如Hormonal Contraceptives, IUD…。在這個model並沒有如XGBoost所認為性行為的年齡是一個重要的因素，或者說在我們的認知中過早的性行為會影響Dx:Cancer，在XGBoost中可以印證我們的想法，但在CatBoost中因為有更多的資訊，發現有其他比初次性行為更加重要的因素，如IUD（宮內節育器），或許暗示著我們或許會忽略，但他對癌症所帶來的影響實際上不容小覷。

三.特徵分群分析

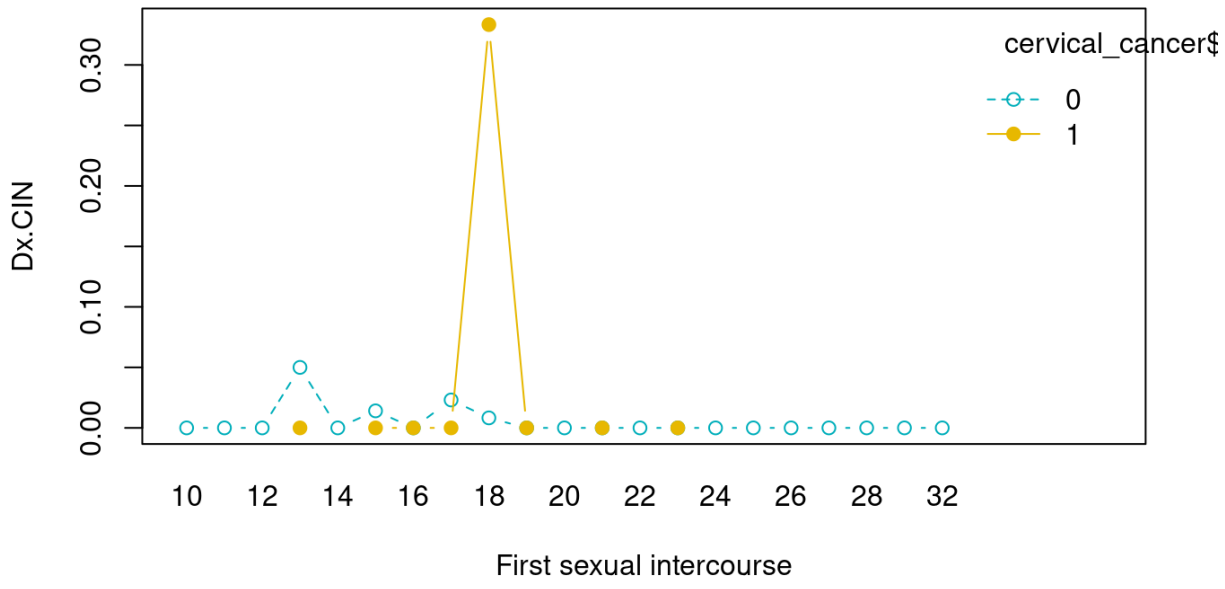
特徵劃分:Feature Clustering

我們希望可以把36種不同類型的feature進行分群，我們規劃變數有以下的群:基本個資，行為個資，疾病相關的生理變數與診斷的變數。以下是我們認為個變數所屬最適合的群。

- 基本個資 :Age, Number of sexual partners, Num of pregnancies
- 行為個資 :First sexual intercourse (age), Smokes, Smokes (years), Smokes (packs/year), Hormonal Contraceptives, Hormonal Contraceptives (years), IUD, IUD (years)
- 與疾病相關的生理變數 :STDs, STDs (number), STDs:condylomatosis, STDs:cervical condylomatosis, STDs:vaginal condylomatosis, STDs:vulvo-perineal condylomatosis, STDs:syphilis, STDs:pelvic inflammatory disease, STDs:genital herpes, STDs:molluscum contagiosum, STDs:AIDS, STDs:HIV, STDs:Hepatitis B, STDs:HPV, STDs: Number of diagnosis
- 診斷的變數 :Dx:Cancer, Dx:CIN, Dx:HPV, Dx, Hinselmann: target variable, Schiller: target variable, Cytology: target variable, Biopsy: target variable

變數間的交互作用：Interaction between varivbles

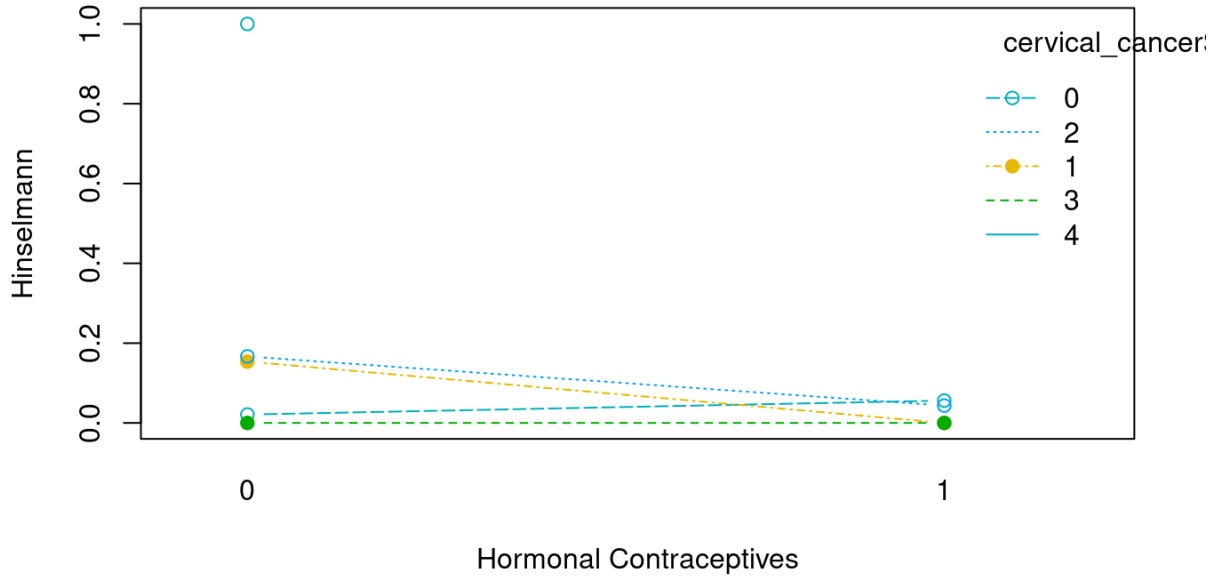
1. 從生活習慣到檢查出的病因中，我們發現了第一次性行為的時間跟HIV－病毒有相當的關聯。



由上面的 Interaction 圖我們可以看出，在18 歲間進行第一次性行為且有愛滋病可能會導致子宮頸癌的機率大幅提升，由於我們資料的特性（得癌症的患者並不多），所以18歲是一個相對重要的年紀。我們分析了幾個原因，

1) 在調查的階段，受到調查的人員因為各種因素可能會謊報其第一次性行為的時間，導致我們資料集中18歲的人佔了非常大的部分。
2) 在18之前的性行為中，保護措施做的比較好，由於我們知道子宮頸癌是透過病毒傳染，如果有做好保護措施的話相對的比較不容易感染。但如果有好好的使用安全措施的話，就也比較不會得子宮頸癌。

2. 接著我們透過分析使用激素性避孕藥跟性病的交互作用，如下圖



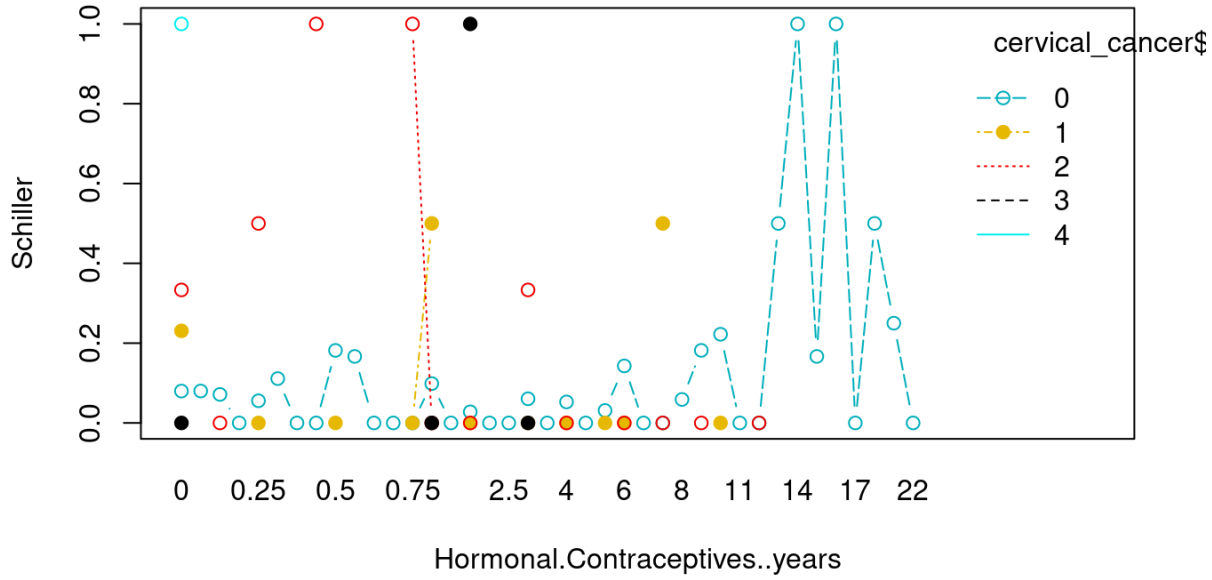
Unset

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Hormonal.Contraceptives	1	0.04	0.0419	0.944	0.331536	
STDs..number.	1	0.17	0.1694	3.819	0.051044	.
Hormonal.Contraceptives:STDs..number.	1	0.48	0.4847	10.926	0.000994	***
Residuals	736	32.65	0.0444			

在較少的性病數量中，在有使用避孕藥的情況下較不容易檢查出子宮內膜病變(這是一種導致子宮頸癌的重要因素)，但如果同時有許多性病的情況下，使用避孕藥較容易使得子宮內膜病變，需要特別注意。

透過這個分析，我們還可以看到對較少性病的人使用避孕藥可以降低子宮內膜出血的情形，但相反的對較多性病的人來說，使用避孕藥可能導致風險提升。

3. 透過更近一步的分析，我們想看看使用避孕藥的時間與性病的交互作用，希望可以找出對子宮頸癌中的另一個重要檢查(細胞染色分析)的結果。

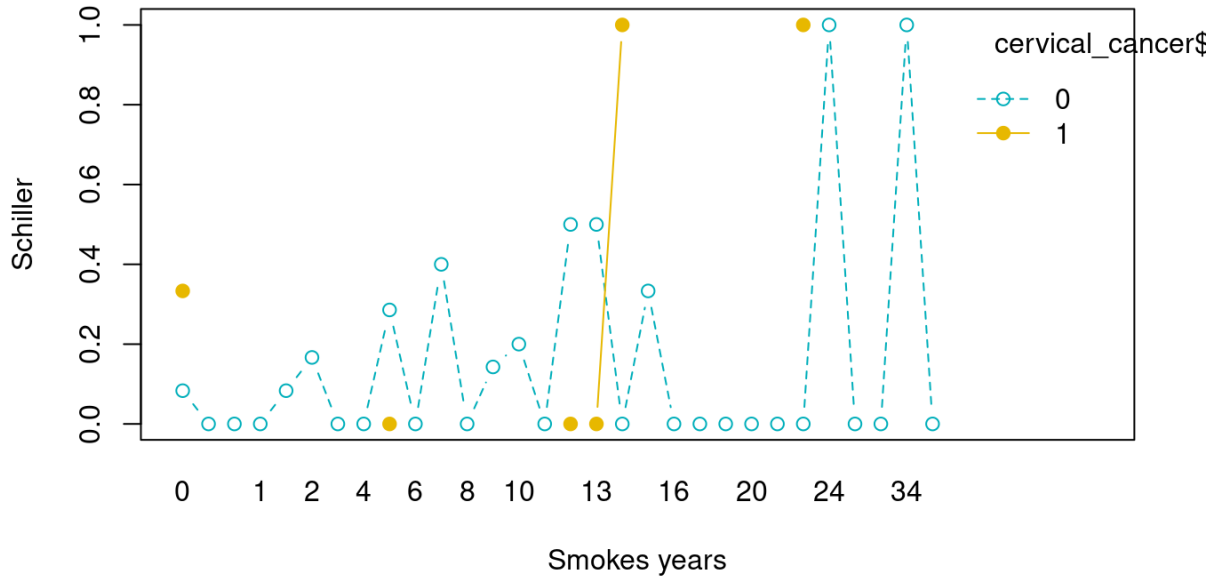



```
Unset
[1] "Schiller"
                                                Df Sum Sq Mean Sq F value  Pr(>F)
Hormonal.Contraceptives..years.                  1   0.53  0.5262   6.117 0.01362 *
STDs..number.                                    1   1.09  1.0881  12.647 0.00040 ***
Hormonal.Contraceptives..years.:STDs..number.    1   0.87  0.8660  10.066 0.00157 **
```

從上面的圖我們可以發現到在低年限的避孕藥使用中，擁有較多性病的患者更有機會檢測出細胞病變，我們發現了幾件有趣的事。

1) 如果你沒有任何性病的話，長期服用避孕藥（10年）以上的話，有更高的機率會導致子宮頸癌的，因此我們應該要注意服用避孕藥的時間。
2) 如果有性病的話，短時間服用避孕藥就有可能造成子宮頸癌機率的上升，這或許是未來服用避孕藥時需要注意的。

4. 接下來我們想看看吸菸與子宮頸癌的關聯，首先我們先用一個吸煙年齡跟HIV的交互作用開始。




```
Unset
                    Df Sum Sq Mean Sq F value   Pr(>F)
Smokes..years.       1   0.23  0.2331   3.660 0.056123 .
STDs.HIV             1   0.80  0.7953  12.488 0.000435 ***
Smokes..years.:STDs.HIV  1   0.27  0.2708   4.252 0.039544 *
Residuals          739  47.06  0.0637
```

在上圖中，我們可以發現到並沒有明顯的趨勢可以解釋說吸菸與 HIV 之間的關聯，但 Two-way anova 的結果卻顯示其有interaction，我想這是因為我們資料的特性所導致的結果。

肆. 討論與結論

首先, 我們先透過將所有的變數放到 logistic regression 中來看什麼是重要的變數, 但我們遇到了變數過多且不平均資料導了模型可能會overfitting, 所以我們透過兩種方法來進行變數選擇, 一種是透過 AIC 來進行 backward selection 另一種是透過 T-test 來進行 backward selection 。透過上述兩種方法, 我們皆得到了比 full model 特異度及敏感度更好的模型, 但仍然存在一些問題如顯著性不佳 , 因此我們使用了機器學習的方法希望能夠得到更好的結果。

以機器學習的方法時作時, 我們使用了兩種不同的分類器, XGBoost 及 CatBoost , 這兩種分氣都可以產生一個決策樹來解釋模型的結果, 除了幫助我們更好的與傳統 logistic model 相互對照, 也是一種解釋力比較強的機器學習方法。透過機器學習的方法 , 我們發現到一些重要的變數如 Dx.HPV 這也是我們已知會導致子宮頸癌的重要原因, 但也發現了一些不同的重要解釋變數如避孕藥的使用, 及子宮內避孕器的影響。

為了更近一步的觀察變數間的交互作用, 我們使用了 Two-way anova 及 interaction graph 來觀察是否某些因子的交互作用會導致子宮頸癌的發生, 透過大量的 anova test 及刪除一些本來就有相關的變數, 我們發現了一些有趣了交互作用如使用激素行避孕藥及性病的交互作用, 透過這些交互作用的發現可以提供相關領與的專家進行近一步的研究。

透過上面的方法解釋了我們想預測的目標Dx. Cancer(子宮頸癌)並得到了良好的特異度及敏感度, 其中 accuracy 達到了 0.9775 , 這些可以方便地幫助我們預測子宮頸癌的發生, 並且解釋子宮頸癌發生的原因, 且找出更多的潛在因子, 能幫助未來子宮頸癌的診斷及風險評估。