

國立政治大學統計學系

碩士學位論文

結合 spline 及分箱方式之廣義線性模型預測

Generalized linear model prediction combined with

spline and binning method



指導教授：黃子銘博士

研究生：楊翔宇撰

中華民國 110 年 6 月

摘要

在日常生活中，總是要面臨許多資料。大部分的資料都是夾雜著類別型變數以及連續型變數的資料。針對這種資料，提出了一個方式可以對自變數稍作些許處理，並以處理後的自變數加以預測資料，達到不錯的效果。

本研究方法將會使用R語言以針對銀行信用卡違約付款的資料作為主要的研究對象。以下個月是否有違約行為作為反應變數，其反應變數以1(有違約行為)、0(無違約行為)做為表示。利用模型可以從中了解信用卡用戶的基本訊息影響違約行為與否的機率，供以衡量信用卡用戶未來將會違約的機率，以幫助銀行對這些客戶進行限制，以降低銀行虧損的風險。

[關鍵字] B-spline、無母數方法、變數選取、分段多項式、節點選取、WOE of binning、分箱方法。

Abstract

In our daily lives, we always have to face a great amount of large datasets. Most of them are combined with categorical variables and continuous variables. Regarding this type of data, we proposed a method for model construction and prediction.

The proposed method is applied to the data of bank credit card default payments as the main research object. The response variable is the payment situation in the following months. “1” means the user with breach of contract and “0” means without breach of contract. Using the model, we can understand the association between the basic information of credit card users and their default behavior, which can be used to measure the probabilities that credit card users will default in the future, so as to help banks monitor customers and reduce the risk of bank losses.

[Keywords] B-spline, nonparametric method, piecewise polynomial, variable selection, knot selection, WOE of binning, binning method.

目 錄

1 緒論(背景介紹)	6
2 文獻探討	7
3 研究方法	9
3.1 連續型變數之處理	10
3.2 離散型變數之處理	16
3.2.1 分箱方法的介紹	17
3.2.2 使用R進行分箱	19
3.3 結合所有變數以獲取最終模型	20
3.4 隨機森林法	23
4 模擬實驗和結果	25
5 實際資料應用	27
6 參考文獻	44

圖 目 錄

3.1	流程圖	10
3.2	分箱結果1	20
3.3	分箱結果2(x4的分段結果)	21
5.1	數值型變數之相關圖	29
5.2	需做PCA項的相關圖	31
5.3	資料中各類別型變數的IV值	34
5.4	資料中類別型變數分箱結果1	35
5.5	資料中類別型變數分箱結果2	35
5.6	資料中類別型變數分箱結果3	36
5.7	隨機森林法中各變數重要性	38



表 目 錄

3.1	預測能力標準表	18
3.2	變數X4分箱區間之對應統計量	19
3.3	分箱後各變數對應的區間(因子化)-以模擬資料中前6筆為例	22
3.4	變數重要性的排序	23
4.1	模擬結果之MSE比較表	26
5.1	資料的變數介紹	28
5.2	各數值型變數之最佳節點	30
5.3	X1~X6的主成分分析	32
5.4	X1~X6的主成分Loadings	32
5.5	X1.1~X6.1的主成分分析	32
5.6	X1.1~X6.1的主成分Loadings	32
5.7	各類別型變數挑出的指標變數個數	37
5.8	MeanDecreaseGini表	39
5.9	類別型變數對應行數表	40
5.10	類別型變數移除及對應BIC	40
5.11	指標意義介紹	41
5.12	訓練集採論文方法配適之三指標	41
5.13	訓練集採隨機森林法配適之三指標	42
5.14	測試集採論文方法預測之三指標	42
5.15	隨機森林法預測機率結果	42
5.16	測試集採隨機森林法預測之三指標	43

第1章 緒論(背景介紹)

分類在機器學習中一直扮演極為重要的角色，常用的分類工具有很多，例如：分類樹法、隨機森林法、支持向量機(SVM)及boost法等...。其中有些方法提供挑選變數的方式，例如：在隨機森林法(random forest)中，會以Mean Decrease Gini值為變數排序重要性，在結合了變數重要性的資訊後，隨機森林法將會以各變數重要性作為依據，讓整個分類模型的分類結果更好。

廣義線性模型(generalized linear model)也是一種常用的分類工具，而相較於隨機森林法，在此提出一套自帶變數選取功能，判斷出哪些變數對於分類為100%有用，讓模型可以對反應變數有著良好的解釋能力及預測能力，並利用這模型做出良好的分類。

本論文主要的貢獻在於針對廣義線性模型提出一個選取變數的方法。資料中的自變數中通常可分為連續型和離散型兩類變數。對於連續型變數，先以spline方式對變數進行轉換，再做變數的挑選；至於離散型變數，則採取WOE(weight of Evidence)方法進行分箱，並以其對於離散型變數分箱的標準，對資料本身原離散型變數改成數個指標化變數的組合，此方法也考量到IV(information value)，結合其觀念對離散型變數的選取更加精確。經由上述所提的變數選取方法處理，將可改善廣義線性模型的解釋能力及預測能力。在論文中，上述廣義線性選取方法對廣義線性模型的分類結果與利用隨機森林法得到的分類結果進行比較，以評估此兩種方法的效果。

第2章 文獻探討

在機器學習的領域中，監督式學習的分類法是採取由訓練集資料中的自變數X及反應變數Y所構成的一個學習模式，並利用此模式以推測或應用於新的資料。模式的輸出或為連續型的結果，亦或是離散型的分類。此篇論文中，主要針對輸出結果為分類的資料進行模型的配適及預測。在現實中，時常也會遇到分類問題，所以很多方法都因此而產生，接下來將針對一些分類問題的研究做出回顧。

- 論文方法模型解析

在此階段，將深入的介紹論文中所提到的各種方法。而論文中所做的事便是應用這些方法，稍加改良後進行調整配適出一個預測力不錯的廣義線性模型。

首先，在此篇論文中，主要是基於廣義線性模型的背景下進行推廣的。廣義線性模型 (generalized linear model)方程式利用連結函數 (link function) 建立各種不同尺度的反應變數與解釋變數間的迴歸。在此篇論文中，利用連結函數的logit的轉換以處理反應變數Y為二項分佈的資料。

第二步，利用spline模型，選取連續型的變數的節點，以構建更完善的模型。Spline函數是一種分段為光滑之方程式，利用多段的分段低次方(論文中設定的函數為3次方)多項式，以保持在分段處具有一定光滑性的函數插值。詳細計算方式可以參考[2]，其中分段的那幾個關鍵點，就是此篇論文所提到的節點，良好的節點選取將提升模型的解釋能力及預測能力，並搭配引用資料時所引入的邊界條件以構成整個spline插值模型。其相對於高次多項式插值可能出現的振盪現象，利用spline模型所構成的函數存在著較好的數值收斂性和穩定性。

第三步，便是搭配WOE分箱(woe of binning)模型，將離散型變數利用分箱的標準進行類別化。根據[10]，早在1994年即有WOE的文獻出現，如[3][7].WOE的字面意思即為證據權重(weight of Evidence)其對分箱後的每組進行計算，這是一種

監督式學習的方法，WOE越大時，代表該分箱具要較高的預測能力去預測反應變數為何。同時，此分箱方法通常會搭配IV值(information value)作為篩選變量的指標，這個指標越大時則意味著該自變數的預測能力越強，類似於信息增益指標的概念，有關於信息增益可參考[9]。而詳細的內容在論文中的研究方法都有更深一步的探討。

- 隨機森林法

除了論文中主要使用的廣義線性模型外，在處理分類問題時，時常也會使用隨機森林模型。隨機森林模型是一種監督式學習模型，最早由Ho[5]提出，其主要是從決策樹模型所衍伸的。

而決策樹(decision tree)是一種呈現樹狀的分類器，可參考[1]，其大致為三個步驟：特徵選擇、決策樹生成以及剪枝決策樹。在特徵選擇的階段，會決定該用哪些特徵做判斷。在訓練集的資料中存在許多變數，而各變數的作用大小不一。故特徵選擇的作用便是篩選出和分類結果具有較高相關性的特徵。並在選擇好特徵後，從根部節點開始觸發，並對節點進行計算所有特徵的信息增益(information gain)。會選具有信息增益最大的特徵當作節點特徵，並根據該特徵的不同取值標準建立子節點，並用同樣的原理產生新的子節點，直到信息增益很小或者沒有特徵可以選擇為止。最後便是主動去掉部分的分枝以降低模型過擬合的風險。

而隨機森林法的基本原理是利用Bagging(Bootstrap Aggregation)的方式抽取決策樹。Bagging就是採取抽取後放回的方法從 n 筆資料的訓練集中抽出n個樣本產生一棵決策樹，並重複該步驟很多次以產生許多棵CART樹（使用GINI算法的決策樹）。它的決策結果是匯總所有不同決策樹的共同投票採多數決所構成的，透過此方式可大幅度的增進最終的運算結果。這種方法也被稱為集成方法(Ensemble Method)。此法的優點主要是可以處理大量的輸入變數及評估各變數的重要性，並且一般來說對於資料的預測都有一定水準的準確度。

第3章 研究方法

本研究中採用的logistic model，型式如下：

$$\log \left(\frac{P(Y = 1|X)}{P(Y = 0|X)} \right) = f(X)$$

式子中Y為二元的反應變數、X為自變數向量以及 $f(X)$ 為X的函數。

在此，考慮到變數向量X中可分為連續變數 $X_{con,1}, \dots, X_{con,p}$ 和離散變數 $X_{dis,1}, \dots, X_{dis,k}$ 兩種，所以先考慮較簡化的可加性模型，如下：

$$\log \left(\frac{P(Y = 1|X)}{P(Y = 0|X)} \right) = f_1(X_{con,1}) + \dots + f_p(X_{con,p}) + g_1(X_{dis,1}) + \dots + g_k(X_{dis,k}), \quad (3.1)$$

其中的 f_1, \dots, f_p 為 $[0,1]$ 區間上的spline函數，若變數不落入此區間內，將對其進行標準化。而 g_1, \dots, g_k 的形式如下：

$$g_j(X_{dis,j}) = \sum_{\ell=1}^{n_j} \beta_{j,\ell} I(X_{dis,j} \in (a_{j,\ell}, b_{j,\ell}]), \quad j = 1, \dots, k,$$

其中 $I(X_{dis,j} \in (a_{j,\ell}, b_{j,\ell}))$ 為指標函數，即 $X_{dis,j}$ 落於區間 $(a_{j,\ell}, b_{j,\ell}]$ 內為1，反之為0； $\beta_{j,\ell}$ 為指標函數的係數。區間的下界 $a_{j,\ell}$ 、上界 $b_{j,\ell}$ 和 n_j 則由分箱結果做決定。當Spline函數的基底決定後，式3.1所包含的參數可由最大概似估計法(Maximum Likelihood Estimation)以進行估計。

接下來的研究方法主要以4節依序介紹內容，分別為：第3.1節將介紹Spline函數針對連續型變數的處理，包含節點的選取以及logistic模型的構建。第3.2節將介紹離散型變數的處理，包含分箱方法及離散型變數重要性排序。第3.3節將介紹當模型中的連續變數及離散變數皆經過處理之後，如何將其結果整合為最終模型。第3.4節則是介紹要與論文方法相比較的隨機森林法之模擬流程。由於文章將介紹的研究方法之細節較多，在

此先繪製研究方法的主要流程，如圖3.1：

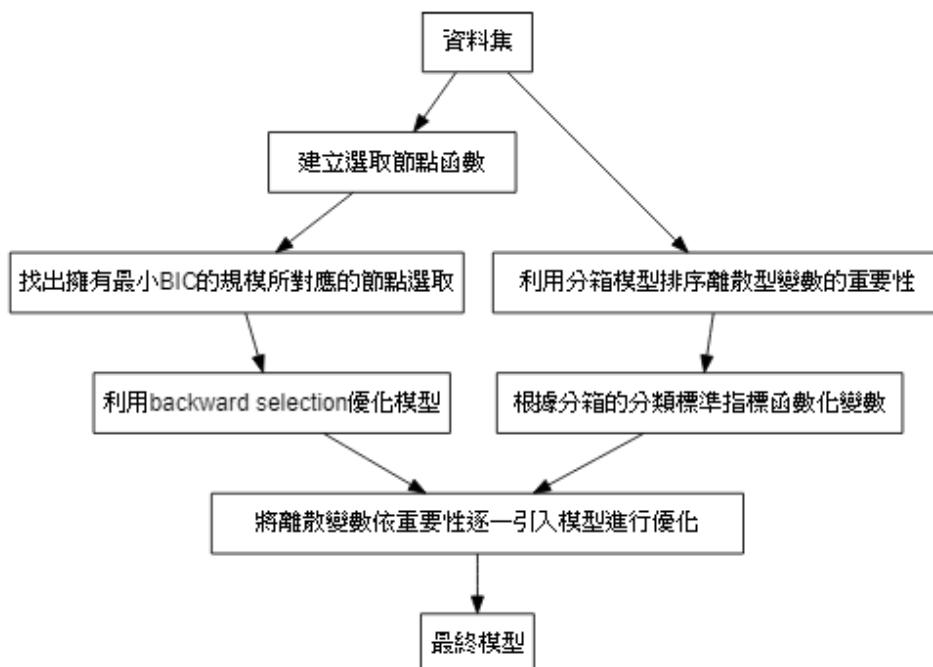


圖 3.1: 流程圖

第 3.1 節 連續型變數之處理

資料中的自變數可分為離散型變數 $X_{dis,1}, \dots, X_{dis,k}$ 及連續型變數 $X_{con,1}, \dots, X_{con,p}$ 兩大類。為便於說明，假設資料中連續型變數分別為 X_1 和 X_2 ，並存在離散型變數為 X_3 至 X_7 以配合模擬實驗的設定。

本節中針對連續變數的處理分為六個步驟，在步驟一到步驟四處理連續型變數時，是針對單一個變數表示進行個別的處理，因此以 X_1 做為範例說明。(選取節點的流程參照黃子銘教授的論文)

- 步驟一：產生待選取節點與對應的檢定統計量

此步驟主要在創造出所有待選取節點並計算檢定統計量，在步驟二時會需要利用此檢定統計量挑選具影響力的節點。根據[6]中所敘述的方法進行函式的編寫，詳細程式請見附錄第3節。

針對單一的連續型變數，做以下的處理：

1. 首先寫出spline中選取節點的函數 knot_selection()，其中此函式有兩輸入值為new_demo1和j.new_demo1為包含單項的連續型變數的資料集，其中包含變數 X_1 以及對應到的反應變數y、而j是指節點所切的規模(scale)，以0.5的j次方作為標準，是自己所設定的，在之後的檢定會用到此參數。

2. 對自變數 X_1 的值由小到大進行排序，接著找出最小值，方式為每個 X_1 的值減掉其中的最小值之後，再除以全距，可使這些標準化後的 X_1 皆落於[0,1]的區間。

3. 再對經標準化後的 X_1 進行唯一化，即取出其中相異值以作為可能選取的節點向量，表示為 $X_{1,i}$,其中 $i = 1, \dots, q$.這會影響到接下來選取節點的標準。

4. 將經唯一化後的標準化 X_1 的值逐一帶入迴圈以求出對應的檢定統計量：

進行第i次迴圈的步驟如下：

先令出第i個唯一化後的 X_1 對應的正指標集x.pos以及負指標集x.neg.

正指標集x.pos定義為 $\{i^*: X_1 \text{的第 } i^* \text{ 筆觀察值落於開區間 } S_i\}$,

$$S_i = (\text{第 } i \text{ 個唯一化後的 } X_1, \text{ 第 } i \text{ 個唯一化後的 } X_1 + (0.5)^j)$$

負指標集 $x.neg$ 定義為 $\{i^*: X_1 \text{的第 } i^* \text{ 筆觀察值落於開區間 } S_i\}$,

$$S_i = (\text{第 } i \text{ 個唯一化後的 } X_1 - (0.5)^j, \text{ 第 } i \text{ 個唯一化後的 } X_1)$$

接著迴圈中進行下面的步驟:

a) 如果以第*i*個唯一化後的 X_1 所構成的正指標集或者負指標集所包含的個數小於20筆時，則不考慮將該 X_1 作為節點，而對下一個進行迴圈測試。

b) 分別對將這些負指標集所對應的 X_1 中的點建立線性迴歸module_neg，及對正指標集所對應的 X_1 中的點建立線性迴歸module_pos.

※注意：若反應變數y皆為0或皆為1時，則無法建立線性函數去配適的，應進行下一個迴圈的測試。同時也需要判斷：模型module_neg或module_pos中若係數估計結果為na的情形，則以進行下一個迴圈的測試。

c) 將分別算出負指標集以及正指標集線性迴歸的均方誤差MSE (Mean-Square Error) .此MSE在R語言中可以透過summary(模型)的cov.unscaled矩陣去乘以誤差變異數的估計。算出兩變異數矩陣的估計後，因為正指標集中對應的資料與負指標集中對應的資料皆為獨立，故接下來要做正指標集和負指標集的迴歸係數是否有差異的檢定時，可直接將此兩變異數矩陣估計相加，得到結果令其為BV.

d) 除了BV之外，還需要將迴歸模型module_neg和module_pos的斜率項係數以及截距項分別進行相減，得到 2×1 矩陣結果令其為a。利用a和BV進行檢定統計量的推算，檢定統計量為：

$$a^T (BV)^{-1} a$$

將此檢定統計量值及對應到的為待選取的節點加到一資料集，當所有唯一化的變數 X_1 均依上述迴圈處理後，接續著以下的步驟。

- 步驟二:挑出符合條件的情況下較顯著的節點

將步驟一得到的待選取節點，依照其檢定統計量值由大到小進行排序後，並作以下操作：

1. 刪除檢定統計量值中不超過臨界值的待選節點。此處的臨界值為卡方分配(自由度為2)的第95百分位數：5.991465。統計量值大於臨界值的節點構成初始的待選取節點清單S.

在R程式寫作上，需要注意以下狀況:

- a) 若待選清單S只有一行時，該資料會直接變成numeric形式，則應將其轉換成行數為2的矩陣才有辦法進行之後的選點步驟。
 - b) 若待選清單S為空清單時，應回傳NULL以代表未選到任何節點，方便進行之後有關spline函數的建構。
2. 不然，待選清單S的行數大於1的情況下，則可將S中的節點位置及對應的統計量值，以及想設的節點規模(scale) k 帶入choose_knot()函數，以尋得具標準化形式之節點。接下來，將探討choose_knot函數是如何構成的。詳細程式請見附錄第3節：
 - a) 初始化待選取節點清單，S
 - b) 從待選取節點清單中，挑出統計量值為最大的節點，令其為 g^* 。
 - c) 將 $g^* \pm (0.5)^k$ 範圍內的節點由待選取節點清單中進行移除。
 - d) 重複b、c直到清單都被取完為止。

其中要注意的是：

若經反覆移除後，待選清單僅剩1個節點，則會直接變成numeric形式，需將其轉換成行數為2的矩陣才有辦法進行選點。不然，當待選清單行數為空清單時，即為選取完畢。

- 步驟三

此步驟主要為建立BIC函數以計算BIC的指標，並和AIC指標比較優劣。透過步驟一和步驟二將有辦法選出連續型變數根據指定尺度所選擇到的節點，接著要透過步驟三選擇究竟是何種尺度，對於模型的解釋例有較好的表現，在此採用的是貝葉斯資訊準則（Bayesian Information Criterion，BIC）原則進行選取。

BIC是根據AIC（Akaike Information Criterion）所改善的指標：

考慮到在訓練模型時，常會增加樣本數量，讓概似函數(likelihood)增大，但此舉卻會導致過擬合的現象。針對此問題，AIC及BIC均使用了與模型樣本個數相關的懲罰項，其中又以BIC的懲罰項比AIC的大，這使得面臨樣本數量過多的情況下，可有效防止模型精度過高所造成的過擬合情形。

BIC指標是由模型的log likelihood、樣本數n以及參數個數k所構成的，其式子如下：

$$BIC = -2\log\text{-Likelihood} + k \log(n)$$

建構好BIC函數之後，接下來要進行步驟四的迴圈測試。

- 步驟四:根據BIC選出最佳規模scale

在節點的選擇時，一般不會讓節點的數量過多導致過擬合的狀況發生，於是在規模(scale)的選擇上將範圍鎖定在3~8之間，探討在何種規模下所產生的模型存在著最小BIC,經過比較後以選擇出其規模(scale)。

在固定的Scale下，計算BIC的步驟如下(詳細程式可參照附錄第3節)：

1. 針對連續型變數 X_1 進行節點的還原，若發現選取節點之後的結果是回傳NULL的，還原後的節點也就為NULL。不然一般而言，由於選出的節點皆為標準化後的結果，所以得將最選到的結果乘以未標準化的 X_1 之全距(range)並加上該變數中的最小值以進行還原，以獲得 X_1 的節點。
2. 此步為給定條件下算出B-spline基底。B-spline基底是利用R語言中splines的套件包中的bs()函數計算，此函數輸入的設定如下：取值點設為原資料的連續變數 X_1 的觀察值、spline的degree設為3、基底的節點設為選到的節點以及節點範圍的界限設為連續變數中的最小值至連續變數中的最大值。其中，參數degree=3是預設分段多項式的次數為三次。
3. 利用所算出的B-spline基底，作為廣義線性模型的解釋變數，而反應變數為資料中的y，並將截距項去除(此舉是為了排除共線性的問題)，以配適模型。

分別將Scale=3~8之下的模型及資料筆數套入步驟三設定好的BIC函數得其結果，並將其存起來，最後選擇出具最小BIC的規模為何。

● 步驟五

此步驟主要為加總各連續變數經B-spline基底轉換後的結果。藉由步驟四比較BIC後所得到的結果，針對每一個變數，根據BIC去挑取具最小BIC的規模，並重新根據該規模進行節點選取以及再將變數進行B-spline基底轉換以得到新的變數。

接下來，要將多個轉換後的新變數進行加總。其中要注意的是，除了第一個連續變數所轉換的新變數之外，其餘的變數所構成的新變數皆須捨棄一個變數，以避免線性相依發生係數無法估計的情況。

最後經B-spline基底轉換的新變數，作為廣義線性模型的解釋變數，反應變數為資

料中的y，並要去除截距項，以產生模型。

- 步驟六

此步驟將利用backward selection 做最終連續型參數選取。在步驟五時，反應變數和連續型變數所構成模型大致上已經完成，而此步驟是為了要將模型優化所設置的。詳細程式請見附錄第3節。

以下為優化的過程：

1. 先把資料中的反應變數y、連續變數 X_1 和 X_2 經B-spline基底轉換後的新變數進行合併，以得到新的資料集w2。
2. 再利用資料集w2進行廣義線性模型的配適，接著利用stats套件包中的step的功能，進行backward selection，逐一剔除一些顯著性不高的變數，以獲得反應變數和連續型變數所構成模型。

使用R語言優化的過程中，若沒有進行cbind()產生新的資料集去做變數選取的話，會因為個別變數經B-spline基底轉換後的新變數會被視為一體，而導致在剔除參數時，會直接將 X_1 或 X_2 的轉換出的新變數整組刪除，而無法進行單一新變數的剔除。

第3.2節 總離散型變數之處理

處理完資料中的連續型變數後，從此步驟開始將進行對於離散變數的處理。首先，先將資料中的離散變數 X_3 、 X_4 、 X_5 、 X_6 和 X_7 以及反應變數y進行合併，另存為一資料集data.discrete。進而利用此資料集進行方法進行分箱。以下是此分箱方法的介紹(見第3.2.1節)及如何使用R進行分箱(見第3.2.2節)：

第3.2.1節 分箱方法的介紹

參考[11]，在此分箱方法中，是以卡方分配和卡方值為理論基礎，藉此觀察某個因素是否對反應變數有所影響。其對應到的卡方檢定之虛無假設 H_0 為該離散變數與反應變數無關，基於此假設下所算出卡方值代表著觀測值與理論值之間的偏離程度，若偏離越大的話，代表著該離散變數與反應變數有關。此卡方檢定也有著一種衡量預測值與觀察次數之間的差異到底有多少概率是由隨機因素引起的，若此概率(似p-value概念)很小時，即為該離散變數在不同狀態之下，反應變數的分配不盡相同。

利用上述所提到的卡方分配及卡方檢驗，利用Kerber提出的ChiMerge法[8]作為核心原理。此原理採取自下而上不斷合併的方法以完成分箱操作。而在進行每一步的合併過程中，依照最小的卡方值找尋最優的合併項。其主張為：若某兩區間可以被合併，則該兩個區間的卡方值是最小的。故ChiMerge的步驟如下(以下例子為順序尺度資料)：

1. 將離散變數A經排序後分成區間較多的若干組。
2. 將鄰近兩組合併後，計算合併組合的卡方值。
3. 找出上一步的卡方值中最小的一個，假設為第*i*−1組與第*i*組所構成的組合，將其合併成新的一組。
4. 重複2、3，直到達成以下情形之一則完成合併：
 - 合併後，最小卡方值超過預設卡方臨界值為止
 - 合併後，區間數達到指定的數目（例如5, 10, 15）

將區間合併後，結合WOE (weight of evidence) 以及IV (information value) 的觀念進行分箱評估。

WOE為證據權重，即對自變數的一種監督式的編碼形式，第*i*組的WOE公式如下：

IV值	預測能力
[0,0.02)	無預測能力
[0.02,0.1)	弱
[0.1,0.3)	中等
[0.3,+\infty)	強

表 3.1: 預測能力標準表

$$WOE_i = \ln(p_{y_i}/p_{n_i}) = \ln\left(\frac{y_i/yes}{n_i/no}\right) = \ln\left(\frac{y_i/n_i}{yes/no}\right)$$

其中， p_{y_i} 為第*i*組中反應變數為1者佔所有樣本中反應變為1的比例、 p_{n_i} 是第*i*組中反應變數為0佔所有樣本中反應變數為0的比例。其意涵為第*i*組中反應變數為1和反應變數為0取對數後的比值，與所有樣本中比值取對數後的差異。

而IV是結合WOE的觀念所推衍的產物，對於第*i*組，IV計算公式如下：

$$IV_i = (p_{y_i} - p_{n_i})WOE_i$$

並將該離散變數的每個分組的IV值進行相加，即可獲得整個變數的IV值。如下：

$$IV = \sum_i^n IV_i$$

而計算出的IV值可以顯現該變數的預測能力為何，本文採取一般IV值的判斷標準(如表3.1)作為判斷。在此會以IV值作為衡量預測能力的指標是因為它有著以下特性：

- IV值是非負的，適合衡量一個變數的預測能力。
- IV值可表現出當前分組的情況下，個體的數量佔整體數量的比例，反映出對於變數預測能力的影響。

不過在使用WOE以及IV觀念時，須注意以下情形：

WOE Table for data.x4									
Final.Bin	T.Count	T.Distr.	1.Count	0.Count	1.Distr.	0.Distr.	0.Rate	WOE	IV
≤ 1	3288	32.9%	1852	1436	29.1%	39.5%	43.7%	-30.5	0.032
≤ 2	1713	17.1%	1712	1	26.9%	0.0%	0.1%	688.6	1.851
≤ ∞	4999	50.0%	2799	2200	44.0%	60.5%	44.0%	-31.9	0.053
Total	10000	100.0%	6363	3637	100.0%	100.0%	36.4%	NA	1.935

表 3.2: 變數X4分箱區間之對應統計量

1. 若最大箱裡面數據的數量佔總數據90%以上，則棄用該變數。
2. 若遇到兩變數相關性高而只須保留一個的情況下，則選擇IV值較高或者分箱較均衡的變數。
3. 若遇到該變數分組所對應到的反應變數皆為1或皆為0時，則須將此分組視為一個規則，作為模型的前置條件或補充條件，並重新對變數進行分組。

第 3.2.2 節 使用R進行分箱

利用R中woeBinning套件包進行分箱，其中可採取woe.binning()函數進行自動分箱，裡面須填入離散變數的資料集data.discrete以及資料中的反應變數y(須以括號包含表示)。如下：

```
binning <- woe.binning(data.discrete, 'y', data.discrete)
```

在進行自動分箱後，利用woe.binning.table()函數(裡面須填入上一步分箱的結果binning)，可將分箱後的解決方案根據各變數進行表格儲存。如表3.2:

利用woe.binning.deploy()函數可將分箱解決方案binning部署及應用到新的數據集，但此處將結果引用原數據集data.discrete，如下：

```
df.with.binned.vars.added <- woe.binning.deploy(data.discrete, binning,
```

```
add.woe.or.dum.var = 'woe',min.iv.total = 0.00001)
```

此處設定中的add.woe.or.dum.var = 'woe'是使在新增變數時帶有WOE分數，設定min.iv.total = 0.00001是使最小顯現分箱結果的IV為0.00001。部署後的結果可看到各變數的分箱以及對應到的分數。同時利用woe.binning.plot(binning)，可將分箱後結果可視化，得到圖3.2和圖3.3：

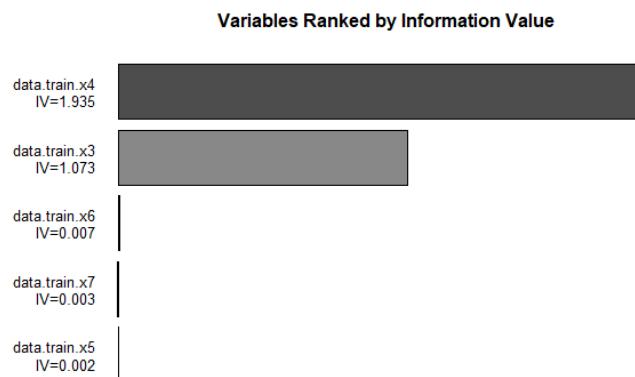


圖 3.2: 分箱結果1

這裡將儲存圖3.3的分箱區間以及圖3.2的變數IV值，接下來將根據這些結果，將離散變數添加在先前第3.1節已完成的模型之上，讓模型更加完善。

第 3.3 節 結合所有變數以獲取最終模型

這裡的連續變數將以經過backward_selection 後的模型作為基準，以下程式碼是從原本的資料集w2中取出backward_selection 選好的部分(儲存於back_result2)。如下：

```
a = colnames(w2)  
b = rownames(summary(back_result2)$coefficients)  
p = w2[a %in% b]
```

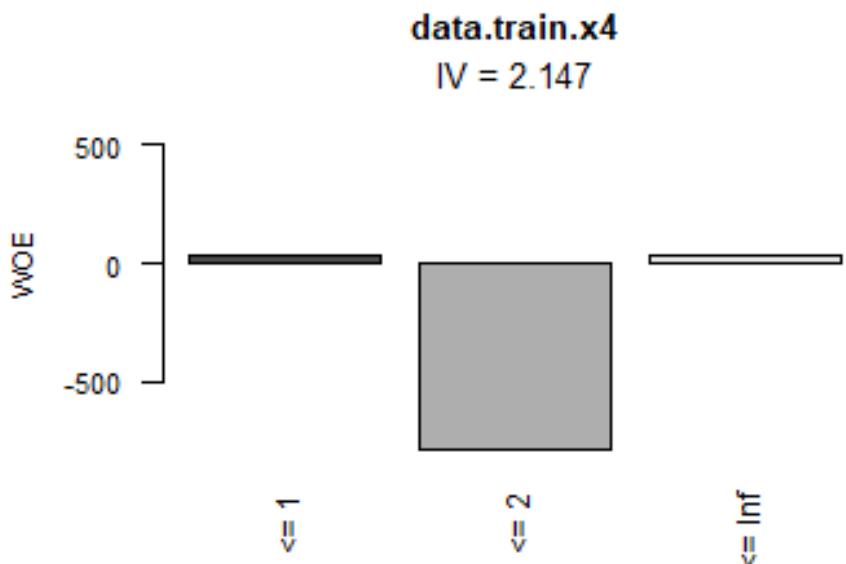


圖 3.3: 分箱結果2(x4的分段結果)

```
V1 = w2$V1
w2 = data.frame(cbind(V1,p))
```

上方程式碼中，先依序令a為原w2的行名、b為經過backward selection後的模型仍被保留下來的變數，再利用p表示那些原資料集w2中的自變數依舊在b裡面依舊存在的變數究竟有哪些。在此反應變數y以V1表示，因為先前進行cbind()時，反應變數名稱被自動改成V1。因為p並未包含反應變數，故需將原資料集w2中的反應變數V1額外取出與p進行合併，作為之後被當作基準的模型資料，並再次命名為w2。

下一步，須利用到R語言中fastDummies和recipes套件包，將在先前步驟所進行分箱的各變數根據所對應的區間轉換為因子(factor)形式，並將其存取作為之後代入迴圈優化的資料para_sim(已按照IV由大至小排序)，如表3.3型式。

接下來，利用dummy_cols()函數，將各變數根據對應的區間轉換成指標函數，亦即對於某變數時，若符合該區間值為1，否則為0，此步驟可較清晰判斷出離散資

name	x4	x3	x6	x7	x5
1	(2, Inf]	(1, Inf]	(3, Inf]	(-Inf,3]	(0,2]
2	(-Inf,1]	(1, Inf]	(-Inf,2]	(-Inf,3]	(-Inf,0]
3	(2, Inf]	(1, Inf]	(-Inf,2]	(3, Inf]	(4, Inf]
4	(2, Inf]	(1, Inf]	(2,3]	(-Inf,3]	(2,3]
5	(2, Inf]	(1, Inf]	(-Inf,2]	(-Inf,3]	(0,2]
6	(-Inf,1]	(1, Inf]	(-Inf,2]	(-Inf,3]	(0,2]

表 3.3: 分箱後各變數對應的區間(因子化)-以模擬資料中前6筆為例

料位於某特定區間有著較顯著的特徵。同時也要注意：在利用此函數時，須將參數remove_selected_columns設為TRUE以蓋掉原先變數，方便進行迴圈。除此之外，若先前分箱部署的資料得到的因子有缺失值，而原始資料沒有時，則須將該因子剔除。

將離散型變數轉換成因子後，接下來要把這些因子加入模型中，一次加入一個離散變數的對應指標函數並作調整，指標函數加入的順序是依照IV值大小，由大至小逐一加入，最後依照模型BIC大小進行挑選模型。

模型比較的步驟如下：

1. 令model_ok為利用資料集w2所配適的廣義線性模型。
2. 在尚未被考慮加入模型的離散型變數中，選出IV值最大的離散型變數，將其對應的指標函數加入model_ok中。接著配適廣義線性模型，並且逐一刪除最不顯著的指標函數，直到所有指標函數皆為顯著($p\text{-value} < 0.05$)為止，將所得到的模型稱為model_challenge。
3. 將model_challenge及model_ok所對應到的BIC值進行比較。若model_challenge的BIC值小於model_ok的BIC值時，則將model_ok更新為model_challenge，否則，則不更新model_ok。
4. 重複2.和3.直到沒有未考慮過的離散型變數。
5. 最後以model_ok為所選的最終模型。

變數名稱	MeanDecreaseGini
x6	296.5868
x7	302.0160
x5	306.3227
x3	702.3061
x4	748.1868
x2	943.1708
x1	1160.6990

表 3.4: 變數重要性的排序

第 3.4 節 隨機森林法

為了呈現本論文所採取統計方法的效能，選擇和隨機森林(random forest)法做比較。

利用同筆資料，搭配上R語言中randomForest套件包進行以下操作：

1. 先設定模型m1為使用randomForest()函數所生成的模型，裡面先填入反應變數及自變數和所使用的資料。
2. 利用模型m1預測出給定各資料的自變數情況下，反應變數為0及為1的機率各自為何。並利用varImpPlot()函數，觀察模型中各自變數的Mean Decrease Gini值，並可透過此值以衡量變數的重要性，其意義為 Gini 係數減少的平均值。原理是：在隨機森林中，衡量變數的重要性方法乃採取剔除該變數，並將剔除變數後的模型與原模型比較，如果之間差異越大則表示該變數越重要。
3. 依Mean Decrease Gini值由小到大，將所對應到的變數進行排序，如表3.4。同時，需將那些離散變數 X_3 、 X_4 、 X_5 、 X_6 和 X_7 取出，作為之後選取變數的標準。
4. 從m1模型中，將變數依變數的重要性由低到高一一剔除，並分別構成不同模型。
5. 利用getTree()函數，算出各模型中節點個數為何，其中要把split var為NA項進行過濾，則可得到模型的參數個數。同時，也須將利用m1預測出的條件機率進行

以下的相乘：若真實 $y=1$ 者，取對應到預測 $y=1$ 的機率；若真實 $y=0$ 者，則取對應到預測 $y=0$ 的機率，得到likelihood，以求取BIC。另外，在進行模擬實驗時，要用R語言randomforest()模型算出的 $Y = 1$ 之條件機率 P^* ，並將 P^* 與實際 $Y = 1$ 的條件機率 P_0 相減，求取MSE，其中

$$P_0 = P(Y = 1 | \text{真實模型中所有的} X_i).$$

此處MSE的定義為模擬中所有 P^* 和 P_0 相減的均方和去除以模擬次數。而將採取其最小的BIC及其對應到的MSE作為與論文中統計方法比較的模型。



第4章 模擬實驗和結果

針對第3章所提出的研究方法，進行了一個模擬以進行實驗，以比較論文方法和隨機森林法之表現。

首先在自變數生成上，模擬數據包括連續型和離散型自變數，連續型變數為 X_1 、 X_2 ，分布為uniform(0,1)，離散型變數為 X_3 、 X_4 、 X_5 、 X_6 和 X_7 ，是從0、1、2、3、4、5中等機率抽樣生成，在此每個變數生成10000筆觀察值，以得到所需的自變數資料。

根據上方所生成的自變數資料，生成反應變數 Y 的資料。生成的方式是根據羅吉斯模型(logistic regression)，即

$$L = \ln \left(\frac{P(Y = 1|X_1, X_2, X_3, X_4)}{1 - P(Y = 1|X_1, X_2, X_3, X_4)} \right), \quad (4.1)$$

其中

$$L = 3X_1 - 2X_2^2 + 5 * I(X_3 = 1) + 9 * I(X_4 = 2) - 1.$$

上式中， $I(X_i = j)$ 為指標函數值，當 $X_i = j$ 時為1，否則即為0。此處 $i \in \{3, 4\}$, $j \in \{1, 2\}$

為了方便生成0、1這種類別變數，需要了解 $P(Y = 1|X_1, X_2, X_3, X_4)$ 為多少，利用(4.1)反推可以得知 $P(Y = 1|X_1, X_2, X_3, X_4) = e^L / (1 + e^L)$ ，利用此機率進行二項分配抽取10000次亂數，生成對應的0和1的反應變數，並將此變數進行因子化(factor)，得到反應變數 Y 的10000筆觀察值。

以上為生成一次資料的流程，產生好資料以後將使用第三章研究方法建立模型，並且獲得 $Y = 1$ 的條件機率估計。接著需將重複以上流程500次，以獲得500次的MSE。針對隨機森林的方法，也會用同樣的500筆資料進行機率的估計。

在以MSE作為衡量模型方法表現的情況下，500次的模擬中，論文中的統計方法皆小於隨機森林法，代表論文中的統計方法較佳。表4.1為各方法500次MSE的平均及標準

MSE比較表		
統計方法	平均	標準差
論文方法	0.006412	0.0097389
隨機森林	0.046042	0.0018357

表 4.1: 模擬結果之MSE比較表

差。

在變數選取上的話，論文方法中，真正重要的離散變數 X_3 、 X_4 皆有被選取到。除此之外還會選取其他離散變數，隨機森林法中，被選來做比較的模型時常只選取到一個重要的離散型變數，另一個重要的離散變數，無法被選取到。所以在模擬的階段中，無論是以MSE或是變數選取做標準，皆為論文方法較佳。

而在模擬中，為了要存取模型的表現為何，須存取各離散變數提供模型多少指標變數，以及存取利用model.ok所配適出的條件機率與先前生成數據的條件機率 $P(Y = 1|X_1, X_2, X_3, X_4)$ 的MSE，如此操作即完成一次此統計方法的模擬。因資料為10000筆、2個連續自變數、5個離散自變數以及1個離散反應變數，一次模擬需花費約8分鐘的時間成本。在此，筆者為進行模擬，利用平行運算函數parLapply()，裡面須將上面所敘述的研究方法寫成函數帶入平行運算函數，此舉可節省大量時間。

第5章 實際資料應用

- 資料為KAGGLE中所提供的資料集，是台灣2005年4月到9月的30000筆有關信用卡帳單支付及背景資料
- 資料結構為30000筆資料，25個變數

以下是變數介紹，見表5.1

- 資料處理
 1. 將資料中的類別型變數之數值進行因子(factor)化。
 2. 資料中的類別型變數PAY_0、PAY_2、PAY_3、PAY_4、PAY_5、PAY_6包含了-1、-2等等的負數值，在此這些負數值一律設成0作為無拖延付款，並將以上的變數因子化。
 3. 由於BILL_AMT1~BILL_AMT6為2005年9月至4月的帳單金額，因其彼此間具有高度正相關性(如圖5.1)，為避免彼此間具有多重共線性以新變數BILL_AMT代表這六個變數整體的表現，其為BILL_AMT1~BILL_AMT6的平均。
- 切割資料集

為了衡量方法的預測效率，透過R語言設定隨機亂數對30000筆資料隨機抽取21000筆做為訓練集資料(credit.train)，剩餘的9000筆做為測試集資料(credit.test)，接著，對訓練集的資料建構模型以對測試集的資料預測。
- 針對訓練集資料，先對數值型的變數進行處理
 - 先將先前於研究方法中所用以呈現各可能節點和利用該節點所對應到的卡方統計量之函數knot_selection (j,new_demo1)，將其按卡方統計量由大

變數名稱	類型	解釋
ID LIMIT_BAL SEX EDUCATION MARRIAGE AGE	數值型 數值型 類別型 類別型 類別型 數值型	每個客戶的ID 以新台幣為單位的信用額度（含個人和家庭/補充信用額） 性別（1 =男性， 2 =女性） 1 =碩博士， 2 =大學， 3 =高中， 4 =其他， 5 、 6=未知 婚姻狀況（1 =已婚， 2 =單身， 3 =其他） 歲數
PAY_0 PAY_2 PAY_3 PAY_4 PAY_5 PAY_6	類別型	2005年9月的還款狀態 （-1 =無拖延， 1 =延遲一月， …， 9 =延遲九月以上） 2005年8月的還款狀態（與判斷標準與PAT_0相同） 2005年7月的還款狀態（與判斷標準與PAT_0相同） 2005年6月的還款狀態（與判斷標準與PAT_0相同） 2005年5月的還款狀態（與判斷標準與PAT_0相同） 2005年4月的還款狀態（與判斷標準與PAT_0相同）
BILL_AMT1 BILL_AMT2 BILL_AMT3 BILL_AMT4 BILL_AMT5 BILL_AMT6	數值型 數值型 數值型 數值型 數值型 數值型	2005年9月的帳單金額（新台幣） 2005年8月的帳單金額（新台幣） 2005年7月的帳單金額（新台幣） 2005年6月的帳單金額（新台幣） 2005年5月的帳單金額（新台幣） 2005年4月的帳單金額（新台幣）
PAY_AMT1 PAY_AMT2 PAY_AMT3 PAY_AMT4 PAY_AMT5 PAY_AMT6	數值型 數值型 數值型 數值型 數值型 數值型	2005年8月的付款金額（新台幣） 2005年7月的付款金額（新台幣） 2005年6月的付款金額（新台幣） 2005年5月的付款金額（新台幣） 2005年4月的付款金額（新台幣） 2005年3月的付款金額（新台幣）
default.payment .next.month	類別型	2005年10月拖延帳款狀況(1 =是， 0 =否)

表 5.1: 資料的變數介紹

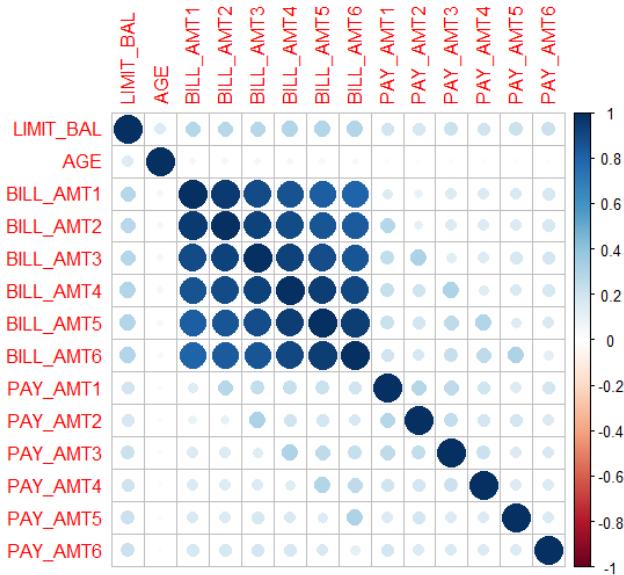


圖 5.1: 數值型變數之相關圖

至小進行排序，並排除掉一些不顯著的節點(即其卡方值低於5.991465)，並搭配選擇適當節點的函數。choose_knot(data_frame,k)，以構成新函數one.button_select (data,k)，以供之後選取節點時方便進行。

- 將感興趣的數值型變數PAY_AMT1~PAY_AMT6及先前創造出的新變數BILL_AMT分別與反應變數(default.payment.next.month)構成一個新的資料集，筆者在此令為A1~A6和B1，如下：

```
response <- credit.train$default.payment.next.month

A1 <- data.frame(PAY_AMT1 = credit.train$PAY_AMT1,y = response)

A2 <- data.frame(PAY_AMT1 = credit.train$PAY_AMT2,y = response)

A3 <- data.frame(PAY_AMT1 = credit.train$PAY_AMT3,y = response)

A4 <- data.frame(PAY_AMT1 = credit.train$PAY_AMT4,y = response)

A5 <- data.frame(PAY_AMT1 = credit.train$PAY_AMT5,y = response)

A6 <- data.frame(PAY_AMT1 = credit.train$PAY_AMT6,y = response)
```

變數名稱	最佳節點規模
PAY_AMT1	5
PAY_AMT2	7
PAY_AMT3	5
PAY_AMT4	5
PAY_AMT5	4
PAY_AMT6	6
BILL_AMT	3

表 5.2: 各數值型變數之最佳節點

```
BI <- data.frame(PAY_AMT1 = credit.train$BILL_AMT,y = response)
```

接著利用上面的這些資料集建構一個大的列表(list) conAMT.list進行選取，並對這些資料集一一代入3~8的規模(scale)，並搭配spline套件包的bs()做B-spline基底轉換，以結合廣義線性模型函數glm()建構模型，並以BIC作為指標，觀察於數值型變數中，何種規模所建構出的模型存在著最小的BIC，詳細程式請見附錄第3節。：

- 選出的各數值變數的最佳節點數如表5.2:
- 透過表5.2所選取的節點數在各數值變數中找出對應的節點，並將這些標準化後的節點進行縮放還原(=選到的節點值 × 該變數的全域 + 該變數的最小值)。
- 並將上面這些還原後的節點依bs()函數進行spline的三次方函數配適。其中要注意，除了PAY_AMT1之外，為了避免線性相依的問題於模型配適中出現，必須得讓其他數值變數進行bs()函數後的結果除去最後一行，以建構之後要進行廣義線性模型配適所需的資料。
- 須額外令新的資料集w，其是由各數值型變數進行spline配適後並結合原始的反應變數(default.payment.next.month)所組合而成的。並要注意：先前已採取因子化的反應變數，若合成新的資料集時需將該值減1，才可進行之後廣

義線性模型的配適。

- 在此發現，經重新建構的資料集w中的前12行(由PAY_AMT1及PAY_AMT2進行spline的結果)存在些許關聯性，如圖5.2：

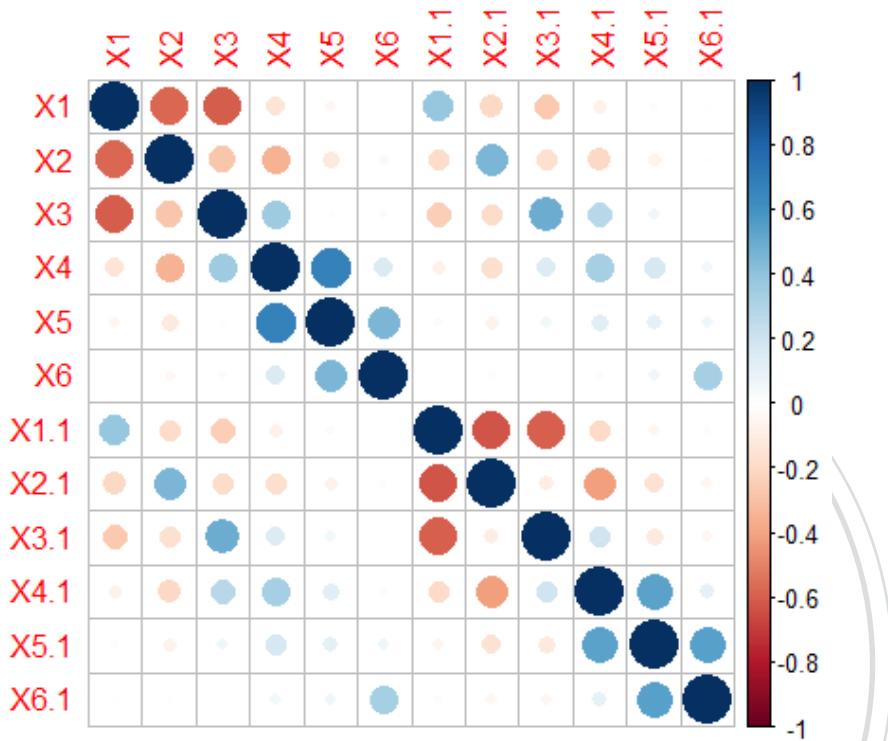


圖 5.2: 需做PCA項的相關圖

根據[4]的想法，故採取主成分分析以進行維度的縮減，使變數更加精簡且更具代表性，以促使模型的估計更精確。

- 將上圖中X1~X6 和 X1.1~X6.1分別以princomp()函數進行主成分分析，結果如表5.3 ~ 5.6

可以看到這兩次的主成分分析(分別令為pca.w1和pca.w2)中，皆在第4個主成分時就有95%以上解釋變異的能力了，故都採取4個主成分進行之後模型的配適，利用主成分分析的loadings對原始變數經B-spline基底的轉換結果做線

Importance of components						
Indicator	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	1.44341	1.30579	1.13477	0.85101	0.446720	1.41365e-07
Proportion of Variance	0.34724	0.28418	0.21462	0.12070	0.033256	3.33067e-15
Cumulative Proportion	0.34724	0.63142	0.84603	0.96674	1.000000	1.000000e+00

表 5.3: X1~X6的主成分分析

Loadings						
Variable	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
X1	0.177	0.736	0.000	0.000	0.000	0.647
X2	0.262	-0.487	-0.561	0.246	0.123	0.551
X3	-0.363	-0.419	0.509	-0.338	-0.220	0.521
X4	-0.607	0.000	0.000	0.428	0.658	0.000
X5	-0.546	0.157	-0.375	0.305	-0.665	0.000
X6	-0.319	0.138	-0.519	-0.741	0.248	0.000

表 5.4: X1~X6的主成分Loadings

Importance of components						
Indicator	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
Standard deviation	1.40058	1.34147	1.12682	0.84745	0.50081	1.18541e-02
Proportion of Variance	0.32694	0.29992	0.21162	0.11970	0.04180	2.34199e-05
Cumulative Proportion	0.32694	0.62686	0.83848	0.95817	0.99998	1.000000e+00

表 5.5: X1.1~X6.1的主成分分析

Loadings						
Variable	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6
X1.1	0.196	0.709	0.000	0.125	0.000	0.657
X2.1	-0.461	-0.293	-0.544	-0.268	0.000	0.574
X3.1	0.000	-0.524	0.486	0.499	0.273	0.402
X4.1	0.506	-0.287	0.308	-0.476	-0.514	0.276
X5.1	0.575	-0.192	-0.323	-0.214	0.694	0.000
X6.1	0.398	-0.133	-0.512	0.625	-0.412	0.000

表 5.6: X1.1~X6.1的主成分Loadings

性組合，以產生新的主成分，操作如下：

```
AMT_1.PCAVAR <- as.matrix(w[,2:7])%*%pca.w1$loadings[,1:4]  
AMT_2.PCAVAR <- as.matrix(w[,8:13])%*%pca.w2$loadings[,1:4]
```

產生新主成分後，取代原始的X1~X6 和 X1.1~X6.1放置於資料集w中。

- 接下來將進行廣義線性模型的配適，並於配適後採取BIC作為標準，以優化模型。以下是根據BIC所採取backward selection的流程：

```
m.bsp.con3<- glm(V1 ~ .-1,family = "binomial",data = w)  
back_result <- stats::step(m.bsp.con3,  
    scope = list(upper=m.bsp.con3),  
    direction="backward",k = log(21000))
```

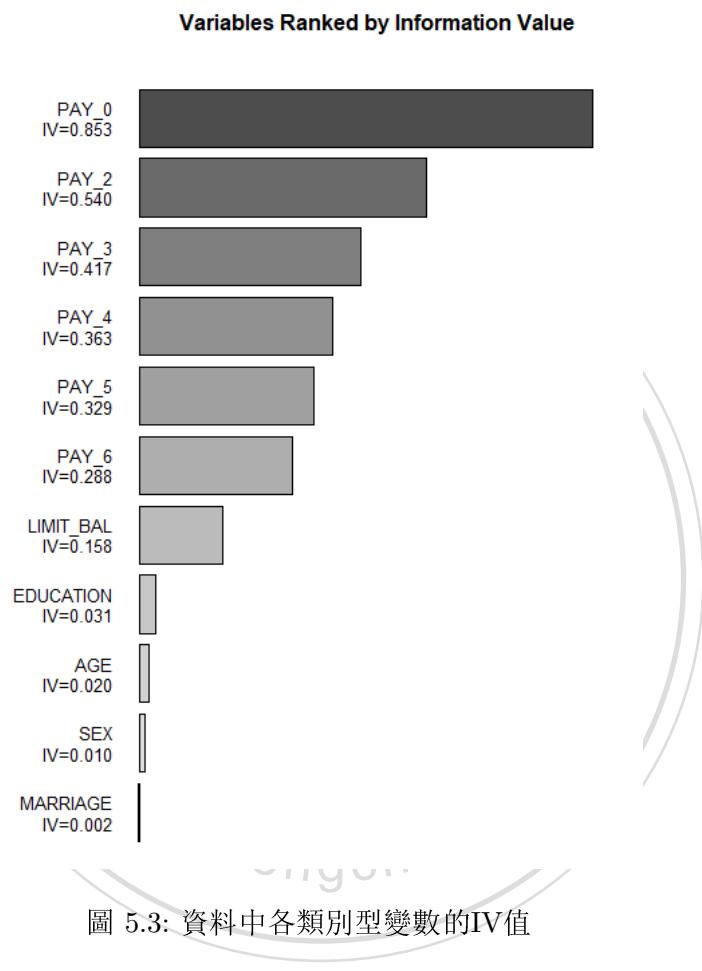
- 針對訓練集資料中的類別型變數進行處理 首先，需先將訓練集中的類別型變數提取出來，令為新的資料集credit_discrete，並利用先前研究方法中敘述過的woe.binning()函數對其進行自動分箱，分箱後按其標準對訓練集資料進行部署。同時，也能觀察到針對訓練集資料中的變數重要性(IV值)及分箱標準與圖5.3~5.6：

接著須引入先前資料集w中最後存在於上一階段的模型之變數資料，其指令如下：

```
w2 = w[ ,colnames(w) %in% rownames(summary(back_result)$coefficients)]
```

上面的指令執行後並未包含反應變數，所以須額外從資料集w中取出反應變數合在新資料集w2裡，以作為之後類別型變數的挑選基準。

下一步就是按照變數重要性(IV值)的排序順序由大到小，一組組代入模型挑選該類別型變數中較顯著的某幾個指標變數並與基準模型比較BIC，比完之後若納



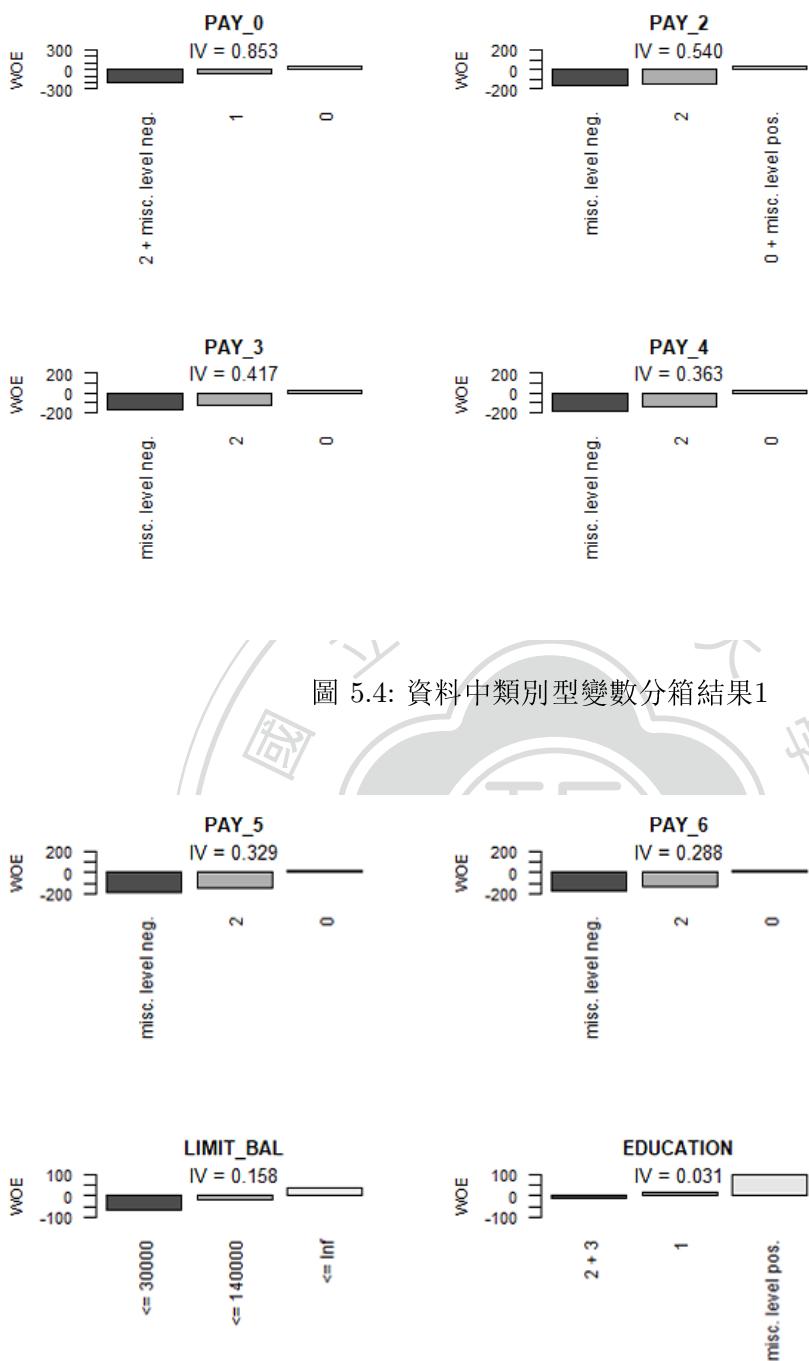


圖 5.4: 資料中類別型變數分箱結果1

圖 5.5: 資料中類別型變數分箱結果2

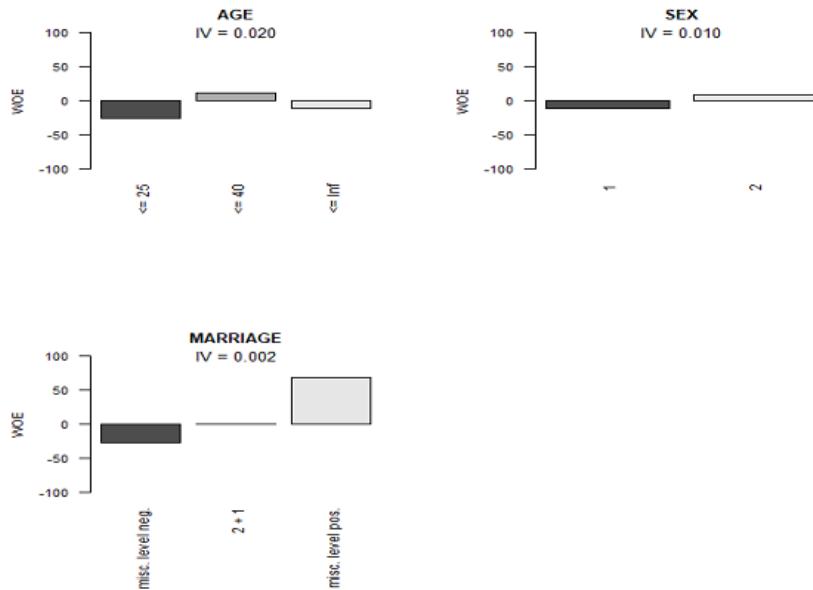


圖 5.6: 資料中類別型變數分箱結果3

入新類別型變數的模型BIC較小，則將該指標變數合成入原基準資料w2以作為下一組變數帶入所比較的基準模型資料。按照此邏輯將資料中11個類別型變數一一代入嘗試，可以看到從各離散變數中挑出的指標變數個數如表5.7：

透過以上的步驟即可完成模型的建構。同時須額外記錄一下被挑選的離散指標變數名稱，以供之後製作測試集資料的變數與訓練集資料的變數對齊。

- 供模型使用的測試集資料製作

- 數值型變數

將原本於訓練集階段各數值型變數所選取的節點，搭配測試集資料(credit.test)進行bs()函數，進行spline的配適所得出資料。配適後，將對應的反應變數與spline配適後的結果進行合併，令其為w.test。

接著利用先前主成分分析所得到的Loadings對測試集資料的相應變數進行線性組合，以獲得測試集資料的PAY_AMT1和PAY_AMT2的各四個主成

變數名稱	個數
PAY_0	2
PAY_2	2
PAY_3	2
PAY_4	1
PAY_5	1
PAY_6	1
LIMIT_BAL	2
AGE	1

表 5.7: 各類別型變數挑出的指標變數個數

分，並將這些主成分取代w.test內原始的PAY_AMT1和PAY_AMT2為資料集w.test.pca。

- 類別型變數

利用先前對於訓練集的分箱標準，在測試集上進行部署，其指令如下：

```
df.with.binned.vars.added <- woe.binning.deploy(credit_discrete,binning,
add.woe.or.dum.var = 'woe',min.iv.total = 0.00001)
```

並將這些分箱後的結果存取，並存取分箱後具因子化的變數令其為para_sim。

再將所有類別型的變數利用dummy_cols()函數進行指標函數化，並只保留在訓練集的最終模型所存取的最後那一些指標變數即可。要注意的是：由於訓練集的資料經過的轉換，導致變數名稱改變到無法和測試及變數名稱對齊。所以須將測試集中對應到離散指標變數名稱改為先前所存取的變數名稱，才可以進行模型的預測。

而將數值型變數及類別型變數的處理的結果進行合併即可成為之後帶入模型預測的資料。預測出的結果會是反應變數的logit，須對其還原成反應變數的值為1的條件機率才可以做準確性的比較，指令如下：

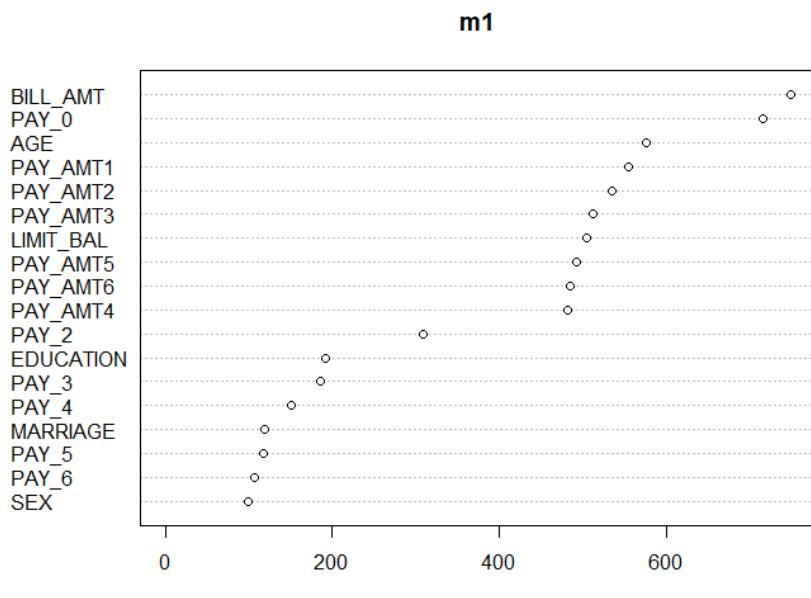


圖 5.7: 隨機森林法中各變數重要性

```

para_sim_xi <- dummy_cols(data.frame(para_sim),remove_selected_columns = TRUE)

para_test_box_select = para_sim_xi[,c(2,3,6,7,10,11,15,18,22,25,26,35)]

colnames(para_test_box_select) = train_boxname_select

test_model_data = cbind(w2.test,para_test_box_select1)

exp_logit = exp(predict(model_ok,test_model_data))

test_predict_prob = exp_logit/(1+exp_logit)

```

<隨機森林法>

而對於同訓練集資料及測試集資料，也採取了隨機森林法與論文中所提出的方
法表現做比較，步驟如下：

- 先初步針對訓練集資料進行randomforest模型的建構，其模型稱為m1。
- 根據模型m1，利用varImpPlot()觀察模型中各自變數的Mean Decrease Gini值，而圖5.7是各自變數的Mean Decrease Gini值排序(由大至小)

Variable	Mean Decrease Gini
SEX	99.46373
PAY_6	106.51656
PAY_5	117.62635
MARRIAGE	119.60806
PAY_4	151.67893
PAY_3	185.82216
EDUCATION	192.45269
PAY_2	308.81025
PAY_AMT4	481.78566
PAY_AMT6	485.27533
PAY_AMT5	493.01597
LIMIT_BAL	505.05320
PAY_AMT3	512.41122
PAY_AMT2	535.40571
PAY_AMT1	554.62582
AGE	575.81270
PAY_0	715.95375
BILL_AMT	749.84988

表 5.8: MeanDecreaseGini表

接著須依Mean Decrease Gini值由小至大進行排序，以供之後模型比較BIC時可以逐一剔除，將會得到表5.8:

- 接著挑出上表內屬於類別型自變數的那11個變數(不含ID)，並找出那11個類別型變數在原始訓練集資料(credit.train)裡所在的行數依序為何，此順序將會用於之後移除資料行以構建新模型進行比較。如表5.9:
- 接下來，將一一移除行並比較模型的BIC，而模型的BIC之計算需要得到參數k的個數及該模型的likelihood。參數k由getTree()函數，算出模型中節點個數為何；而模型的likelihood是採取m1預測出的條件機率進行以下的相乘：若真實y=1者，取對應到預測y=1的機率；若真實y=0者，則取對應到預測y=0的機率。進行比較從一開始全部類別型變數皆納入的情況比較至類別型變數皆不納入，觀察在甚麼階段會有最小BIC。結果如表5.10：

可看到移除前2小的類別型變數，將會得到最小的BIC值，故將其當成最後的隨機森林模型，以供預測測試集資料。

Variable	位置
SEX	2
PAY_6	11
PAY_5	10
MARRIAGE	4
PAY_4	9
PAY_3	8
EDUCATION	3
PAY_2	7
LIMIT_BAL	1
AGE	5
PAY_0	6

表 5.9: 類別型變數對應行數表

變數移除個數	BIC
0	34827.49
1	34396.10
2	33596.35
3	34915.59
4	34103.04
5	34448.47
6	34678.42
7	34506.61
8	34194.54
9	34265.98
10	∞
11	∞

表 5.10: 類別型變數移除及對應BIC

指標	意義	容易導致該指標變低
召回率 (recall)	真正為陽性的樣本中被預測多少陽性樣本的比例	真正為陽性的樣本卻被預測為陰性者極多
準確率 (precision)	所有預測為陽性的樣本中為真正的陽性樣本的比例	預測為陽性的樣本中真實卻為陰性者極多

表 5.11: 指標意義介紹

召回率	準確率	F1-score
0.3421567	0.6844618	0.4562419

表 5.12: 訓練集採論文方法配適之三指標

- 供模型使用的測試集資料製作

針對隨機森林模型，僅需要將測試集資料中如同先前訓練集資料移除前2小的類別型變數即可利用predict()函數預測。

- 訓練集資料配適狀況比較

在比較模型配適之狀況時，將採取常見的F1-score做為衡量模型配適程度的指標。此指標為召回率 (Recall) 及準確率 (Precision)的調和平均數，此兩指標的意義如表5.11：

而針對訓練集資料，利用論文方法之model_ok的配適值(fitted.value)作為依據判斷反應變數(default.payment.next.month)為1或者是0。在此以0.5作為評斷標準，意即：當model_ok的配適值大於0.5時判為下個月將會違約(反應變數=1)；當model_ok的配適值不大於0.5時，則判為下個月將會違約(反應變數=0)。

表5.12為使用此方法各項指標：

而在隨機森林法中，由於每次進行隨機森林法的結果都有細微的不同，故在此是以set.seed(100)固定亂數，以獲得一致的結果。表5.13為使用隨機森林法的各項指標(指標來自於該模型的混淆矩陣)

召回率	準確率	F1-score
0.9648514	0.9982043	0.9812445

表 5.13: 訓練集採隨機森林法配適之三指標

召回率	準確率	F1-score
0.3507647	0.6516957	0.4560616

表 5.14: 測試集採論文方法預測之三指標

- 對測試集資料預測的狀況比較

同樣以召回率、準確率及F1-Score作為衡量對測試集資料預測的能力。

利用論文中的方法所預測出的值會是logit的形式，需將預測出的值先還原成機率的形式，再利用0.5作為分界，以判斷反應變數為1或者是0：當預測出的機率值大於0.5時判為下個月將會違約(反應變數=1)；當預測出的機率值不大於0.5時，則判為下個月將會違約(反應變數=0)。表5.14為使用此判定方式所預測的各項指標：

而利用隨機森林法所預測的結果，會先顯現出預測反應變數為1和0的機率各為多少，如表5.15：

而利用此機率作為判斷反應變數為何，意即：當判1的機率大於判0的機率，就

test_ID	0	1
1	0.476	0.524
3	0.746	0.254
5	0.792	0.208
14	0.558	0.442
16	0.530	0.470
20	0.526	0.474
25	0.788	0.212

表 5.15: 隨機森林法預測機率結果

召回率	準確率	F1-score
0.3621115	0.6594789	0.4675159

表 5.16: 測試集採隨機森林法預測之三指標

判反應變數為1；當判1的機率不大於判0的機率，就判反應變數為0。

表5.16為使用此判定方式所預測的各項指標：

可以看到：隨機森林法在配適模型時F1-score遠高於論文方法，但應用在測試集的預測時，F1-score卻沒有高出許多。這可能是因為用隨機森林法去配適訓練集資料的模型可能有點過擬合的狀況

除此之外，可能是因為自己設定的判定標準和先前在模擬的資料中單純比較預測機率的誤差有著些微差距。所以在預測機率時，可能論文方法會優於隨機森林法，但在需要判定反應變數為1或0時，隨機森林法的表現以些微的差距優於論文方法。

而相較於隨機森林法，由於論文方法是經過一步一步的spline配適並搭配分箱逐漸將模型所配適出的，故透過論文方法的數據是可再現的，不會如同隨機森林法在配適訓練集的模型時，每次的模型都略有差異。

- 改進方向

- 對於應用的資料F1-score有點低，可考慮結合SMOTE過抽樣或者是欠抽樣的方式，對這筆反應變數不平衡的資料(約4:1)做更好的配適。
- 可透過交叉驗證法(Cross-validation)提高配適及預測準確度。
- 在未來可以將此論文所提出的方法應用於透過傳統的線性模型直接用在反應變數為連續型的資料，進行一個不錯的配適及預測。

參 考 文 獻

- [1] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA, 1984.
- [2] C. de Boor. *A Practical Guide to Splines*. Springer Verlag, New York, 1978.
- [3] J. F. Gamble. Asbestos and colon cancer: A weight-of-the-evidence review. *Environmental Health Perspectives*, 102:1038–1050, 1994.
- [4] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [5] T. K. Ho. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1)*, pages 278–282, Montreal, Que., Canada, 1995. IEEE Computer Society.
- [6] T. M Huang. A knot selection algorithm for splines in logistic regression. In *Proceedings of the 2020 3rd International Conference on Mathematics and Statistics*, page 29–33, New York, NY, USA, 2020. Association for Computing Machinery.
- [7] J. Jinot and S. Bayard. Dissent respiratory health effects of passive smoking: Epa's weight-of-evidence analysis. *Journal of Clinical Epidemiology*, 47(4):339–349, 1994.
- [8] R. Kerber. Chimerge: Discretization of numeric attributes. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, AAAI'92, page 123 - 128. AAAI Press, 1992.
- [9] N. Shaltout, M. Elhefnawi, A. Rafea, and A. Moustafa. Information gain as a feature selection method for the efficient classification of influenza based on viral hosts. *Lecture Notes in Engineering and Computer Science*, 1:625–631, 2014.
- [10] D. Weed. Weight of evidence: A review of concept and methods. *Risk analysis : an official publication of the Society for Risk Analysis*, 25:1545–1557, 2005.
- [11] G. Zeng. A necessary condition for a good binning algorithm in credit scoring. *Applied Mathematical Sciences*, Vol. 8:3229–3242, 2014.

附錄

1.Knot selection

```
knot_selection <- function(j,new_demo1){  
  new_demo1 <- new_demo1[order(new_demo1$PAY_AMT1),]  
  tmp <- diff(range(new_demo1$PAY_AMT1))  
  new_demo1$PAY_AMT1 <- (new_demo1$PAY_AMT1-min(new_demo1$PAY_AMT1))/tmp  
  PAY_AMT1.uni <- unique(new_demo1$PAY_AMT1)  
  table_PAYAMT1 = NULL  
  for (i in 1:length(PAY_AMT1.uni)) {  
    cond1_pos <- PAY_AMT1.uni[i]< new_demo1$PAY_AMT1  
    cond2_pos <- new_demo1$PAY_AMT1 < (PAY_AMT1.uni[i]+0.5^j)  
    pos = which(cond1_pos & cond2_pos)  
    x.pos = new_demo1[pos,]  
    cond1_neg <- (PAY_AMT1.uni[i]-0.5^j) < new_demo1$PAY_AMT1  
    cond2_neg <- new_demo1$PAY_AMT1<PAY_AMT1.uni[i]  
    neg = which(cond1_neg & cond2_neg)  
    x.neg = new_demo1[neg,]  
    if (length(pos)<20|length(neg)<20){  
      next  
    }  
    else{  
      if(all(x.neg$y==0)|all(x.neg$y==1)|all(x.pos$y==0)|all(x.pos$y==1)){next}  
      module_neg <- lm(y~ PAY_AMT1,data = x.neg)  
      module_pos <- lm(y~ PAY_AMT1,data = x.pos)  
      ensure_na_neg <- any(is.na(module_neg$coefficients))  
      ensure_na_pos <- any(is.na(module_pos$coefficients))  
      if(ensure_na_neg|ensure_na_pos){next}  
      B_var.neg <- summary(module_neg)$cov.unscaled*(summary(module_neg)$sigma)^2  
      B_var.pos <- summary(module_pos)$cov.unscaled*(summary(module_pos)$sigma)^2  
      BV <- B_var.neg+B_var.pos
```

```

a = module_neg$coefficients-module_pos$coefficients
test_stat <- (t(a) %*% solve(BV) %*% a)[1,1]
column_PAYAMT1 <- c(test_stat,PAY_AMT1.uni[i])
table_PAYAMT1 <- rbind(table_PAYAMT1,column_PAYAMT1)
}
}
return(table_PAYAMT1)
}

```

2. Choose knot

```

choose_knot <- function(data_frame,k){
  s_pos <- data_frame[1,2]+(0.5)^k ;s_neg <- data_frame[1,2]-(0.5)^k
  knot_want <- data_frame[1,2]
  repeat{
    cond = which(data_frame[,2]<s_neg | data_frame[,2]>s_pos)
    data_frame = data_frame[cond,]
    if(is.null(dim(data_frame))){ 
      data_frame = matrix(data_frame,ncol = 2)
    }
    if(length(data_frame[,1])==0){break}else{
      next_knot = data_frame[1,]
      s_pos <- next_knot[2]+(0.5)^k;s_neg <-next_knot[2]-(0.5)^k
      knot_want <- c(knot_want,next_knot[2])
    }
  }
  return(unique(knot_want))
}

```

3.BIC

```
BIC <- function(model,n){  
  k = length(model$coefficients)  
  t = logLik(model)*(-2)+(log(n)*k)  
  return(t)  
}
```

4.根據BIC選出最佳scale

```
for (i in 3:8) {  
  if (is.null(one.button_select(var1,i))){  
    knot_points.var1 <- NULL  
  }else{  
    dif <- diff(range(var1$PAY_AMT1))  
    knot_points.var1 <- one.button_select(var1,i)*dif + min(var1$PAY_AMT1)}  
    PAY_AMT1.bs <- bs(data.train$x1, degree = 3,knots = knot_points.var1,  
    Boundary.knots = c(min(var1$PAY_AMT1),max(var1$PAY_AMT1)),intercept = TRUE)  
    m.bsp.P1_3<- glm(data.train$y ~ PAY_AMT1.bs-1,family = "binomial")  
    Indicator <- BIC(m.bsp.P1_3,10000)[1]  
    Op <- cbind(i,Indicator)  
    X_1 <- rbind(X_1,Op)  
}
```

5.Backward selection

```
w2 <- data.frame(cbind(data.train$y,x1.bs,x2.bs))  
xspline_con2 <- glm(V1 ~ .-1,family = "binomial",data = w2)
```

```

back_result2 <- stats::step(xspline_con2,
                           scope = list(upper=xspline_con2),
                           direction="backward")

base_bic = BIC(back_result2,10000)

```

6. 資料中選擇適當節點規模

```

variable_order = c(19:24,26) #變數行數位置

for (i in 1:7) {
  AMT_1 = NULL
  for (j in 3:8) {
    diff_range = diff(range(conAMT_list[[i]]$PAY_AMT1))
    mini = min(conAMT_list[[i]]$PAY_AMT1)
    knot_points.A1 <- (one.button_select(conAMT_list[[i]],j)*diff_range)+ mini
    PAY_AMT1.bs <- bs(credit.train[,variable_order[i]], degree = 3,knots = knot_points.A1,
    Boundary.knots = c(min(conAMT_list[[i]]$PAY_AMT1),max(conAMT_list[[i]]$PAY_AMT1)),
    intercept = TRUE)
    m.bsp.P1_3<- glm(credit.train$default.payment.next.month ~ PAY_AMT1.bs-1,family = "binomial")
    Indicator <- BIC(m.bsp.P1_3,21000)[1]
    Op <- cbind(j,Indicator)
    AMT_1 <- rbind(AMT_1,Op)
    print(c(i,j))
  }
  spline_list[[i]]=AMT_1
}

```