

學術論著

遺傳規畫決策樹模型於房貸提前償還之風險管理

Genetic Programming Decision Trees in the Risk Management of Mortgage Loan Prepayment

李沃牆* 聶建中**

Wo-Chiang Lee*, Chien-Chung Nieh**

摘要

本文應用無母數的GP-DT模型於房貸提前償還的風險管理，並與文獻上常用的Logit模型比較。實證結果顯示，無母數的GP-DT模型之整體平均正確率均高於Logit模型；而考慮不同的切割值時，Logit模型無論在樣本內外均無一致性，GP-DT模型則呈現一致性的結果，其中以GP-DT500的整體平均正確率最佳。此外，演化代數愈高的GP-DT模型亦有較佳的評估績效。透過誤判成本觀察，我們亦發現，不同的GP-DT模型仍一致性地優於Logit模型。而除Logit-3外，當成本率增加時，各模型的誤判成本卻降低了，其中以GP-DT 500-1、GP-DT 500-2最佳。我們應用敏感性分析來瞭解各特徵變數對提前償還的影響程度，其重要性依次為授信用途、擔保品座落縣市、職業、貸款契約利率、性別、年齡、貸款金額、屋齡、貸款年限。最後，本文亦提出房貸提前償還的風險管理系統，可做為實務上的參考。

關鍵詞：提前償還、遺傳規畫決策樹模型、風險管理、成本率、誤判成本

ABSTRACT

This paper applies a nonparametric GP-DT model to the risk management of mortgage loan prepayment. While previous empirical studies used the Logit model as a benchmark, the results of this study show that GP-DT models have better accuracy on average than Logit models. When considering different cut-off values, the Logit model is not consistent either in sample or out of sample. However, the GP-DT models present consistent conclusions. The GP-DT 500 has a higher overall average accuracy. In addition, the higher the number of generations, the better the GP-DT model's performance. Furthermore, the GP-DT model is also consistently superior to the Logit model in terms of misclassification costs. When the ratio costs increase, the misclassification costs also decrease in most models, except the Logit-3. Among the GP-DT models, the GP-DT 500-1 and GP-DT 500-2 models offer the best performance. Using sensitivity analysis, this study also examines the influence of the key variables on prepayment. In order of importance, these variables include the loan purpose, location, type of occupation, loan interest rate, sex, age, the loan amount, year of housing, and loan period. Finally, this paper also proposes a practical risk management system for mortgage loans for reference.

Key words: prepayment, GP-DT, risk management, ratio cost, misclassification cost

(本文於2008年9月30日收稿，2009年3月24日審查通過，實際出版日期2009年6月)

* 淡江大學財金系副教授。Associate Professor, Department of Banking and Finance, Tamkang University.

** 淡江大學財金系教授。Professor, Department of Banking and Finance, Tamkang University.

聯絡作者：E-mail: wclee@mail.tku.edu.tw; Tel: (02)26215656#3327; Fax: (02)26214755

作者非常感謝二位匿名審查委員對本文提出的許多寶貴意見。

一、緒論

商業銀行大部份的消費性貸款及企業貸款均與信用風險產生高度的相關性，一旦信用風險控管不佳，極有可能面臨虧損甚至倒閉的危機。而銀行風險管理的新規範—新巴塞爾資本協定(New Basel Capital Accord，簡稱Basel II)將銀行所可能面臨的信用風險、市場風險及作業風險，全部納入資本計提的考量範圍內，希望藉由較合於現今環境需要的新架構，進一步來提升銀行風險衡量方式，以期能更有效的規範銀行的風險承擔能力，促進金融體系穩定發展，而此新協定的基本架構主要是由三大支柱組合而成，第一支柱是最低資本要求(minimum capital requirements)，第二支柱為主管機關的監理覆審作業(supervisory review process)，第三支柱則是市場紀律(market discipline)。

長期以來，商業銀行的消費性貸款中(註1)，承作不動產抵押貸款業務一直是相當重要的業務。近年來，包含台灣在內的全球房地產市場交易相當活絡，加上政府提供各種的優惠房貸政策(註2)，為房地產注入一池活水，更助長了金融機構不動產抵押貸款業務。根據央行的統計，截至2008年7月底，本國一般銀行、外商及中小企銀行承做建築貸款餘額已突破新台幣1兆元大關，而銀行辦理的購屋住宅貸款餘額則達到新台幣4兆7,468.27億元。據此可知，銀行對個人或企業放款的額度不僅與日俱增，全體銀行的房地產授信接近滿水位，復以貸款利率的調升，信用風險更是水漲船高。2007年4月2日美國第二大次級房貸公司新世紀金融公司(New Century Financial Corp.)聲請破產，揭開次級房貸風暴序幕，引發全球性的信用市場危機，接著於2008年9月又發生二房危機及美國第四大投資銀行雷曼兄弟控股公司宣佈破產，這場金融風暴至今仍未落幕。因此，金融機構在承作此項業務時，做好風險管理是絕對必要的，因為住屋貸款中最主要的風險即為違約風險(default)與提前償還風險(prepayment)二項，而後者通常是指抵押貸款借款人隨時有權利提前償還部份或全部貸款本金而無須接受任何懲罰(註3)，因為這兩個原因均會造成貸款支付的停止，使現金流量不確定性提高。如何規避這些風險，一則推動不動產抵押債權證券化，加強金融機構資金的流動速率，讓資金從不動產抵押貸款中釋放出來，提高金融機構的資金規劃與使用效用，以免因「借短放長」造成極大的流動性風險。再則是對於提前償還率的預估，往往會受到財務面及非財務面的因素影響，前者如貸款成數、貸款金額等貸款條件屬性；而後者除了住宅屬性、貸款人屬性，亦可能受到購屋者換新工作、離婚或家中成員辭世等因素影響而增加其評估的困難度。

針對「提前還款率」的研究，文獻上已有一些理論模型可資參考，如利用選擇權的觀念來建立提前償還率預測模型，但此模型僅著墨於利率變化和抵押貸款帳面價值的關係。此外，也有應用負債權益來表示債務人違約的賣權價值，即當權益為負時，債務人會考量提前償還，但此類模型似乎忽略了許多極重要的影響變數。

Dunn & McConnel(1981)首度運用隱性有限差分法—CIR利率模型來建構提前償還預警模型(註4)，之後Green & Shoven(1986)提出比例違約模型(proportional hazards model, PHM)，利用鎖住效果(lock-in effect)來探討利率的變動及房屋出售還清(due-on-sale)條款對提前償還率的影響(註5)。Ambrose & Sanders(2001)依據一家保險公司的資料庫，藉由競爭風險模型(competing risk model)來估計商業抵押貸款的提前償還率、違約率及到期的機率。Ambrose & LaCour-Little(2001)則研究採浮動利率的住宅貸款之提前償還率是否低於固定利率。而其他如十二年生

命週期法(12-year life method)，固定提前清償率法(constant prepayment rate, CPR)等經驗法則亦被提及，這些方法雖簡單易用，但缺點是一些可能影響提前償還的因素難以考慮於模型中；另一方面是這些模型大部分只利用過去的經驗對未來的趨勢加以估計，可能不適用於每個地區。

除了上述的理論模型外，透過計量模型來驗證及探討提前償還率的實證相關文獻卻很少，信用資料難以取得應是主要因素。這些文獻所用的模型主要以二分類的多變數Logit模型為主，如國內現有的文獻中，劉展宏、張金鶚(2001)、黃文啟(2002)、黃朝基(2003)；國外的文獻如Clapp et al.(2006)均應用Logit模型探討影響提前償還率的因素以及預測。但吾人以為，提前償還率用線性的Logit模型來預測是否能獲得較佳的預測績效及不錯的解釋，頗值商榷，再者是此模型對提前償還率的預測績效不佳，不易解釋提前償還率的預測是如何形成。另外，此模型的誤差分配設定是否適當，在計量上亦面臨質疑與挑戰，而這些文獻大部分無指標模型可供比較，亦無嚴謹的績效評估模型。黃建智(2004)以卜瓦松迴歸(poisson regression model)來估計比例違約模型假設下影響提前清償與違約變數之參數，研究影響抵押貸款借款人之提前償還與違約因素，程博平(2003)則僅針對影響企業提前償還率因素進行探討，並不能明確地建構提前還款率預測模型。

由上述的文獻可知，過去探討提前償還率模型過於強調利率與提前償還的關係或是考慮線性關係來探討影響提前償還的因素，或者應用Logit模型來估計提前償還機率，但我們知道，影響提前償還的因素絕非利率一項，而Logit模型之優點為可解決自變數非常態的問題、所求得之機率值落於0與1之間、模型可適用於非線性的情況，亦是一種被文獻廣為採用的參數化評估模型，缺點是須經轉換步驟才能求得機率值，在提前償還的判斷較不易解釋。因此，本文思索如何應用一個較具適應性學習能力又可考慮多項相關影響因素，並可根據模型所建構的法則較清楚的解釋提前償還率預測如何形成的方法。所以，本文認為以非線性的無母數遺傳規畫決策樹(Genetic Programming Decision Tree, GP-DT)是一種可以滿足上述條件，亦可進一步嘗試的新方法(註6)，因此，本文首先要透過實證來瞭解無母數的GP-DT是否優於Logit模型，其次要驗證不同的切割值(cut-off value)對不同模型的分類能力是否存在顯著的差異？另就GP-DT而言，當演化代數增加時，對GP-DT的分類績效是否有影響？如何透過敏感性分析來瞭解各特徵變數對提前償還的重要性？完整的房貸風險管理系統如何建構？亦是本文所探討的焦點。

本文第二部分為影響提前償還率因素的探討，緊接著是遺傳規畫決策樹介紹，績效評估模型說明於後，第五部分為實證結果分析，最後則結論與後續研究方向建議。

二、影響提前償還率因素之探討

Campbell & Dietrich(1983)是最早探討還款行為的文獻，該篇論文將還款行為分成正常、遲付、違約及提前償還四種模式，其研究結果發現目前及原始的貸款成數、借款者的支付所得比、失業率、房屋是否為新屋、貸款往來年限與逾期及違約的貸款呈正相關，且合約利率與目前貸款利率之波動與逾期及違約之貸款呈負相關。

單就年齡的影響而言，Quingley(1987)的研究發現借款者年齡愈大，移動性越低。Cunningham & Capone(1990)則發現借款者年齡與提前償還的機率成反比。Archer et al.(1997)發

現借款者年齡越大，抵押貸款契約終止的機率越低。而就利率的影響而論，Kau *et al.*(1990) 主要是透過提前還款選擇權模型來探討，該模型將提前還款視為一種美式買權，借款者隨時可以依貸款未清償之餘額買回貸款契約，在不考慮違約前提下，運用顯性有限差分法來建構最適化提前償還策略模型。實證結果顯示，利率波動率增加對提前償還之價值有正向之效果，顯示利率變動對提前還款有一定程度的影響。

而Smith *et al.*(1996)認為貸款成數、房屋座落地點、利率、貸款金額及利率均與違約行為呈顯著正相關，對提前清償行為也有顯著影響，其中，貸款成數與提前清償行為呈負相關，貸款金額及利率則呈正相關。LaCour-Little(1999)利用貸款服務公司所提供自1997年1月至1998年3月之貸款提前清償資料，以借款人年齡、婚姻狀態、貸款金額、契約利率對市場利率之比率、收入、最初貸款成數、現行貸款成數、種族、地區……等為解釋變數，利用Logit模型進行借款人特徵在預測抵押貸款提前清償檢測，經實證分析得出影響借款人提前清償之顯著因素計有：契約利率對市場利率之比率、貸款金額、貸款成數、現行貸款成數、收入、地區等項目。Clapp *et al.*(2006)利用多變數的Logit模型，並考慮借款人的收入、年齡及信用評分等特性來進一步觀察這些變數的顯著性。

在國內相關文獻中，探討不同因素對提前償還的影響較多，施恩(1994)利用台灣土地銀行之債務人於申請貸款時所選擇的還款方式，探討天真的債務人與聰明的債務人之還款速度，該文利用債務人的特徵變數探討影響債務人還款速度的因素。其結論指出，當利率平穩時，對於天真的債務人，有自用住宅者，會降低其還款速度；新購屋者、住在南部者、薪資收入穩定者、收入越高者，還款速度越快；而對於聰明的債務人，有自用住宅者，會降低其的還款速度；基本放款利率越高時債務人的還款速度越快。郭姿伶(2000)應用中國農民銀行的住宅抵押貸款資料進行實證研究，研究變數包括了性別、貸款金額、貸款利率、貸款溢價、債務人所有的房屋價值、農業貸款、央行優惠利率貸款、是否為首次購屋、本金還付方式、貸款每月攤還額度、貸款期間、貸款年齡、地區等因素。實證方法應用Logit迴歸分析以瞭解影響借款人提前償還與逾期還款的主要影響因素。劉展宏、張金鵬(2001)之研究亦運用Logit模型進行實證，並以台灣土地銀行資料研究對象來分析購屋貸款提前清償行為。結果顯示：借款人選擇提前償還與否的影響因素，確有顯著不同。而決定選擇提前償還抵押貸款行為，則受到婚姻狀況、年齡、教育程度、職業、年收入、屋齡、地區別、建物型態、貸款成數、借款金額、借款期限、契約利率等因素影響。黃文啟(2002)的研究係利用某人壽公司個人住宅抵押貸款資料，並使用Logit模型來分析提前償還住宅貸款的影響因素，在借款者的特性中尋找提前償還率的解釋因子。實證結果顯示員工、性別、契約利率、貸款金額、所在地具有顯著性，且年輕未婚都會女性比較傾向小額抵押房屋貸款之提前償還，表示都會女性較具有理財觀念。

陸文傑(2001)及簡士龍(2003)的研究則側重於利率變動的影響，前者採用樹狀法進行實證研究，並探討不同斜率下的殖利率曲線，抵押貸款證券價格如何受到提前償還參數的影響，結果發現，當殖利率曲線的斜率為負時，因利率變動效果所影響造成的提前償還反而使得折價的抵押貸款證券之價格下跌，而不論在何種殖利率曲線形狀時(正，負或水平)，溢價之抵押貸款證券受到利率效果之影響遠大於時間效果之影響。後者則利用雙變數二項式模型，模擬出固定利率住宅抵押貸款中借款人在還款期間內各時間點的利率與房價水準，配合貸款人的決策模型以及事先假定的事務性服務契約中的各種相關成本與約定(註7)，估算出預期現金流

量後，再利用數值方法決定資產證券化過程中相當關鍵的事務性服務權價值，並且運用情境分析法進一步探討在各種不同的經濟環境變數下，住宅抵押貸款事務性服務權所面臨的價格變動風險。該研究發現，面對利率波動較大的環境下，市場利率下跌的機率越大導致借款人提前清償的機率隨之增加，因而使得固定利率貸款事務性服務權的價值下降。

有部分文獻則探討利率以外其他因素對提前償還的影響，陳健豪(2003)認為影響提前清償與違約之影響因素可區分為利率與非利率兩方面，並進一步從貸款特性出發，分析包含貸款期間、貸款年齡、貸款償還方式等七個變數對提前清償與違約之影響，利用Logit模型進行實證分析，實證結果亦顯示非利率因素對借款人不規則還款行為存在顯著影響。林嘉薇(2004)的研究也得到貸款年齡對提前清償率及違約率的影響型態不同，隨貸款年齡增加，提前清償率成S型曲線增加，而違約率則先遞增再遞減，類似Gamma分配。楊顯爵等(2006)以浮動利率，固定金額本息攤還之房屋修繕貸款者為樣本，並計算個別貸款分期提前還款金額。結果發現利差比、股票指數報酬與波動性對中年紀與高餘額群組的清償率有顯著正相關，利差比有高度路徑相依的關係，而中高年紀與高餘額群組的風險承受能力於五年期間由高變低。

何棟欽(2008)也指出，提前清償可分為內生的提前清償(如利率下跌)與外生提前清償(如換工作)。但要建構提前償還的風險管理系統，必先瞭解影響提前償還率的相關因素。我們於探討上述幾篇相關文獻的重要結論之後，可以得出幾項特徵變數對提前償還的影響方向，以做為本文研究的參考。茲將影響提前償還的因素，區分為貸款人屬性變數、住宅屬性及貸款條件屬性等做為解釋變數，分析如下。

(一) 貸款人屬性變數分析

1. 貸款年齡：理論上，借款者年齡越大(小)，借款者移動性越低(高)，發生提前償還的機率應較低(高)，因此預期借款者年齡與提前償還的機率成反比。
2. 性別：據統計，男性的平均薪資高於女性許多，所以男性的經濟能力也較佳，預期提前清償的機率會高於女性。
3. 婚姻狀況：由於已婚者的家庭支出相較於未婚者高，但婚姻狀況是否與抵押貸款提前償還呈正向或負向的關係，仍有待實證上的研究。
4. 教育程度：教育程度高者較具備責任感與履行契約的能力，對提前清償的影響機率也會越高，二者係成正比關係。
5. 職業：不同的職業，其薪資的差異性頗大，但何種職業別對提前償還率或違約有何影響並不確定，必須根據實際資料進行檢定。
6. 所得：所得越高，借款者償債能力越強，提前償還的機率應較高，預期所得與提前償還的機率成正比。

(二) 住宅屬性變數分析

1. 地區因素：一般認為居住在不同的地方，如都市及郊區會影響提前償還率或違約的機率；就都市地區而言，該區主要以服務業、工業以及製造業為主，而郊區則以農林漁牧等第一級產業為主，因此，都市地區的居民平均所得會高於郊區的居民。所得越高，償付貸款能力自然越高，發生提前償還的機率也越高。

2. 房屋坪數：借款者所購買的房屋面積越大，表示其居住之空間較大，生活品質相對較好，貸款違約機率越小，但對提前清償如何影響則仍不明確。
3. 屋齡：屋齡太高，房價易跌難漲，下一個買方可能會遭遇銀行貸款成數太低的問題，進而影響購屋的意願，因此換屋率大減，預期屋齡與提前清償的機率成反比。

(三) 貸款條件屬性變數分析

1. 首次購屋：通常首次購屋貸款的契約利率會比一般房貸為低，而且申請首次購屋貸款的借款者經濟能力也比申請一般購屋貸款的借款者低，因此預期申請首次購屋貸款者的提前清償的機率較小。
2. 貸款成數：貸款成數越高，表示每個月貸款支出佔所得比例越高，自然使得借款人壓力增加，借款者往後的繳款負擔越大，所以發生提前償還率或違約的機率自然升高。
3. 貸款金額：借款金額大者，表示每月須攤還本息大，也代表抵押房屋的買價較高，一旦借款所得下降時，有可能發生逾期還款情形。
4. 契約利率：借款者與金融機構，在契約上簽訂之貸款利率。契約利率越高，提前償還機率越高，且契約利率亦與違約機率成正比關係。
5. 付款所得比：每月貸款償還額對月收入之比例，即付款所得比，此值與提前償還率成反比，付款所得比越高，會造成借款人負擔愈沈重，因而與違約機率成正比。

三、遺傳規畫決策樹模型

遺傳規畫決策樹是一種由遺傳規畫(genetic programming, GP)和決策樹(decision tree, DT)相互結合的技術，本節首先介紹遺傳規畫演算法，再進一步說明相關的決策樹模型，最後說明模型如何整合。

Koza(1992)針對電腦研究上的一項重點問題，「如何使電腦能在不提供它詳盡程式指令下，自發學習解決問題？」而創造了遺傳規劃(註8)。基本上他是根據遺傳演算法(genetic algorithms, GAs)的機制，將原本以位元串構成的單元，推廣到以程式構成的單元。而每個單元的適應度是由其程式對環境的執行結果而定。所以遺傳規畫就是將傳統遺傳演算法的搜尋位元串狀態空間，推廣到搜尋程式行為空間。而隨著其單元結構由位元串修改成程式，所有相應的遺傳機制也必須作調整，以對應各單元的程式結構，為了讓單元程式可以演化，其中的程式必須被表示為剖析樹結構。

在初始時必須依照環境，決定演化單元程式所需的函式集與元素集，以隨機組合函式集與元素集來產生初始群體的各單元程式。在每個世代中，遺傳規畫的適應度評估必須根據不同環境下執行各單元程式的結果來決定。遺傳規畫已應用在許多工程、科學及財經問題等研究領域，具有相當不錯的績效。

在建構剖析樹結構之前，我們先分析及瞭解所要解決的問題，進而找出會使用到的函式集合，稱之為函式集(function set)，而另外一些會用到的資料，稱作元素集 (terminal set)，函式集可能有幾種不同形態的函式：如算術運算子：+、-、×、÷……等等，數學函數：如Sin、Cos、Exp等等。布林函數：如AND、OR、NOT等等。邏輯運算子：如If-Then-Else等等。迴圈操作子：如Do-Until等等。元素集通常是一般的常數或是變數。另外要注意的是：函

式集和元素集的選擇要滿足下列二個性質：封閉性(closure)及充足性(sufficiency)，前者代表每個函式運算之後的結果一定要落在元素集的定義域中。而充足性表示我們找到的函式集和元素集一定要能解決我們的問題，最好是能有效解決問題方法的最小集合。

而決策樹演算法是屬於資料探勘技術(data mining)中建立分類模式的方法之一，可以將資料依照分割條件自動地分類，再運用歸納方法找出資料來源的規則，建立專家系統並進行預測。決策樹的資料結構如同一顆樹，有節點(internal node)與樹葉(leaf)，其中樹的每一個內部節點代表對應某屬性的測試資料，而每一個分支代表此屬性的一個可能性，例如「是」或「否」，而樹末端的葉節點則代表一個類別或類別屬性。

目前較被廣泛使用的決策樹演算法包括分類與迴歸樹模型(classification and regression tree, CART)、互動式的二分類法(interactive dichotomizer 3, ID3)及C4.5等幾種。CART是由Breiman *et al.*(1984)提出，也是早期決策樹演算法中最受歡迎的一種，其最大的優點是可用來進行分類和建立迴歸樹狀圖，以利決策。此演算法以遞迴的方式不斷的將不同屬性的資料分開，最後將相同屬性的資料區分在一個區塊中，每個區塊中分別利用迴歸的方式來配適不同的統計模型，並以二元樹狀結構圖表示，常用於醫藥、科學、金融等領域的研究。

ID3分類樹由Quinlan(1983)所提出，其主要觀念是當資料中有某項顯著特徵，它就會以此特徵將資料分成兩群，然後這兩群中如果又有一個為特殊特徵，就分為二，以此類推，反覆運作，直到所有同一特徵的資料都在一個類別裡為止。其運作流程是，首先選擇一個最佳的特徵作為根節點，由根節點開始，把所有訓練資料依照此特徵分配到其所屬的分支，如果有某一個分支的訓練資料都已是同一個類別時，則此分支即成為一個樹葉，如此，這個分支的推導也就算完成。同時，這個分支也就是一個分類法則。只要在尚未找到樹葉時，分類的動作就得繼續下去，直至所有資料皆屬同一特徵、同一類別才能宣告完成。這個系統稱為觀念學習系統，是各種演算法中最典型的一個方法。

C4.5是ID3的改良版，其運作的原理與過程皆和ID3相同，只不過C4.5可以處理遺漏和含有連續值的預測因子；此外，C4.5更加入了決策樹的修剪功能以及法則轉換功能，所以此決策樹模型在分類的表現上相當不錯。

由決策樹所分類出來的結果，其正確性與決策樹本身取決資料來源的多寡是有極大的關係。換句話說，資料量的大小往往影響到決策樹判斷的正確性，而這些資料的來源一般是由一些過去的經驗、數據的統計、資訊的累積等所組成。因此，經由龐大資料量所建構的決策樹，其決策結果是相當受到信賴的。進一步言之，決策樹模型能夠歸納出容易解釋的決策法則，計算量相對來說不會很大，也可以處理連續和種類字串並清晰的顯示哪些字串比較重要。但其缺點是對連續性的字串比較難推測。對時間序列的數據，需要很多預先處理的工作，當類別過多時，錯誤率會快速增加。

雖然C4.5決策樹模型可透過訓練樣本之訓練而得到最適的決策樹法則，但當訓練集增加至某一程度時，此模型的運算即會產生困難。而本文所使用的遺傳規畫決策樹模型乃結合遺傳規畫與C4.5決策樹模型，在此整合模型中，遺傳規畫族群內的每一個個體都是一顆C4.5決策樹，而決策樹中的特徵項可做為遺傳規畫的元素集。因為透過遺傳規畫的演化機制，好的決策樹能夠自動產生，再經過多代的演化、訓練及篩選，最後可得出最佳的決策法則，同時亦可解決因訓練集增加，而C4.5決策樹難以使用的問題。此模型如何運作呢？首先是二位元

決策樹(binary decision trees, BDTs)的設計，BDTs是定義由節點(node)所連結的圖形，每一個節點有7個元素，而節點又可區分為終端及非終端的節點，每個節點被定義為，

$$node = \{t, label, P, L, R, C, size\} \dots\dots\dots (1)$$

其中，t代表節點的標記，當節點數目t=0時，即是樹根，label代表終端節點的類別標記，P是父母的標記，對根而言，P=NULL，L及R分別代表左端及右端的子樹(小孩)，對終端節點而言，二者的標記皆為NULL，C是一個二分類且互為對立的集合，相對於非終端節點而言， $n=C[0]$ ， $a=C[1]$ ，而決策法則的形成是按某一特徵值而定，如判斷某特徵變數值是否小或等於a，若結果為「是」，則往左端的子樹方向，若為「否」，則向右端的子樹方向走。對一個終端節點而言，C[i]是第i類訓練樣本的數目，會被分類到節點，而終端節點的類別標記主要由下式決定。

$$C[k] = \max_{\forall i} C[i] \dots\dots\dots (2)$$

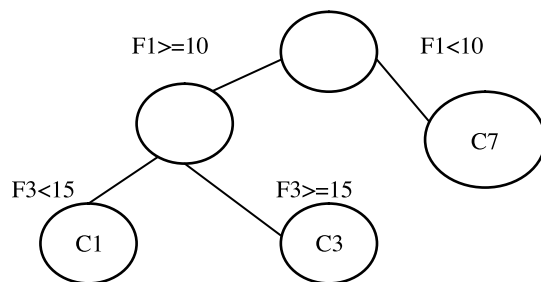
若上式成立則label=k，size是為子樹節點的大小，而根的大小代表整顆樹的大小，終端節點時的大小為1。

建構一個二位元決策樹可經由三個步驟，即分割節點、決定那一個節點是終端節點、分配類別標記至終端節點。應用遺傳規畫來設定一個二位元決策樹，每一個個體可視為一個二位元決策樹。如圖一所示，若F1是首先被檢驗的元素，當小於10，則歸類為C7，若F1大於或等於10且F3小於15，則歸類為C1，否則歸類為C3，而樹的大小為5。

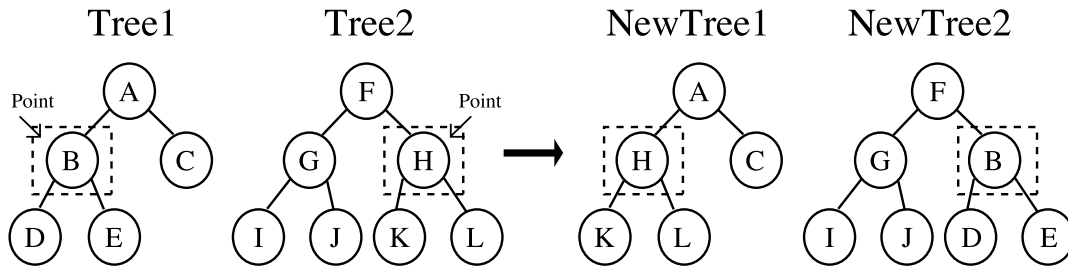
遺傳規畫決策樹的運算子包含選擇(selection)、交配(crossover)及突變(mutation)，如下圖二中的B點與H點互換，即代表決策樹的單點交配(one-point crossover) (註9)。

綜合言之，整個遺傳規畫決策樹的演化流程可用底下二步驟說明，

1. 選擇訓練樣本，然後用C4.5來訓練二位元決策樹模型，反覆進行此訓練，直到每一代族群中所有的樹皆訓練完為止。
2. 用遺傳規畫運作機制來進行決策樹的演化。



圖一 二位元決策樹例子



圖二 決策樹單點交配示例

四、績效評估模型

(一) ROC曲線

ROC曲線是接受者操作特徵曲線(receiver operation characteristic)的簡稱。以往，在信用風險評估模型績效上，常用預測模型的正確率、型一誤差或是型二誤差來呈現，但這些指標是否能充分反映績效？為了釐清這些懷疑並進一步瞭解這些信用風險模型有多少效度(validity)是必要的。一個信用風險模型必須具備一致性或穩定性，也就是所謂的信度(reliability)，再來要求結果是否準確性或精確性，即所謂的效度。效度衡量的方法最常用的CAP (cumulative accuracy profiles)及上述的ROC指標，就二者而言，根據Engelmann *et al.*(2002)的文獻指出，CAP並沒有比ROC提供更多額外的資訊，且CAP的正確率(accuracy ratio, AR)可與ROC的AR做線性轉換(註10)，所以ROC較CAP具實用性。因此，本文僅考慮ROC曲線。

以本文為例，ROC曲線即是在不同的臨界值下，繪製出「模型將正常戶判斷為提前償還戶之數目(型二誤差)」和「模型能正確預測提前償還戶之數目(HR)」之間的關係，如圖三及表一所示。

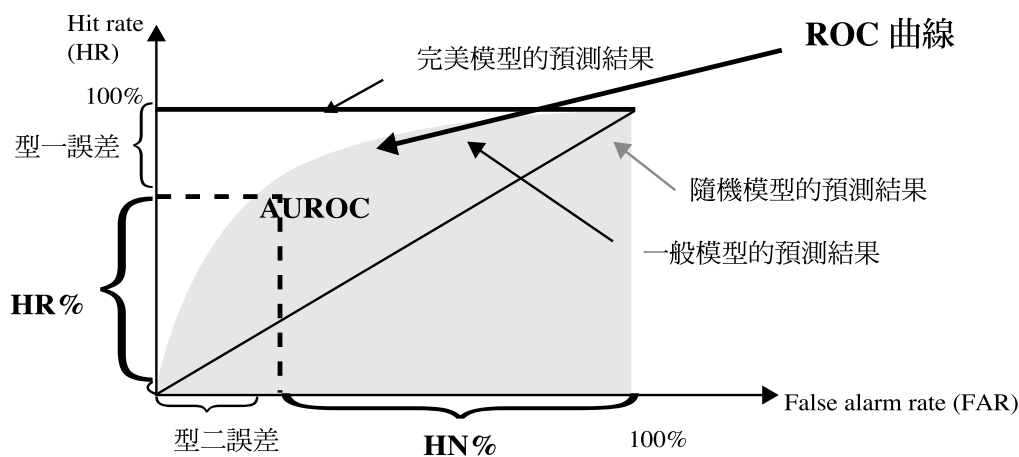
圖三中ROC曲線所包含的面積A即為ROC比率(簡稱AUROC)，Van Deventer & Wang(2003)認為當這比率等於0.5時，表示評分模型是無任何區別力的；若接近1時，則該評分模型為完美的模型，而一般的模型皆介於0.5至1之間，面積的評估公式如下，

$$A = \int_0^1 HR(FAR)d(FAR) \dots\dots\dots (3)$$

另一種判斷方法為在同一個型二誤差之下，型一誤差越小，也表示模型的效力越好。計算公式如下：

$$\begin{aligned} HR(C) &= \frac{H(C)}{N_D} = 1 - P \\ FAR(C) &= \frac{F(C)}{N_{ND}} = P \text{ (型二誤差)} \end{aligned} \dots\dots\dots (4)$$

$H(C)$ ：表切割值為C時，模型能正確預測提前償還戶之數目， $F(C)$ ：表切割值為C時，模型將正常還款戶認定為提前償還戶之數目， N_D ：表提前償還戶數目， N_{ND} ：表正常還款戶數目。



圖三 ROC曲線

表一 給定切割值C的決策結果

	切割值C (門檻點)	實際結果預測結果	提前償還戶	正常還款戶
信用評分	高於切割值	提前償還戶	預測正確(HR)	型二誤差
	低於切割值	正常還款戶	型一誤差	預測正確(HN)

(二) 誤判成本

陳建勝、林明宏(2003)指出，以ROC比率值來衡量模型分類效果時(註11)，必須假設分類錯誤時，其型一誤差與型二誤差所造成的成本率(cost of ratio)等於1(註12)，但如此的假設並不切實際，而就本研究而言，將提前償還戶歸類為正常還款戶，即犯下型一誤差，而對授信銀行房貸風險之影響，超過將正常還款戶歸類為提前償還戶，比犯型二誤差之影響來的嚴重。因此，Boonyanunta & Zeepongsekul(2000)考慮在不同的成本率下，以誤判成本(misclassification cost)來衡量模型因誤判所造成的損失程度，進而比較預測模型的優劣性。成本率會因環境或管理等因素的變化而改變，因此並無固定的數據可供參考。本研究以模擬的成本率來探討其變化所造成的影響，誤判成本的定義如下：

$$\text{誤判成本} = (A + B) / C \dots\dots\dots (5)$$

其中，A = 型一誤差的百分比 × 提前償還戶的比率 × 成本率，B = 型二誤差的百分比 × 正常償還戶的比率，C = 提前償還戶的比率 × 成本率 + 正常償還戶的比率。

整體而言，本研究先以模型在訓練與測試樣本的正確率及整體正確率來比較，再以ROC比率值來觀察不同切割值下的績效。另外，也根據模型的誤判成本來比較Logit模型與GP-DT的優劣性，並檢定兩者是否存在顯著的差異，最後以敏感性分析來說明影響提前償還之重要因素。

四、實證結果與分析

(一) 資料來源及分析

本研究之實證資料取自國內某商業銀行2006年底的購屋貸款案件，並隨機抽取提前清償戶共900筆(不含違約貸款戶)，而正常還款之購屋貸款戶則抽取1,800筆，總計有2,700筆資料。

根據劉展宏、張金鶚(2001)一文對影響房屋提前償還率的分析，將本研究所取得之相關變數區分為貸款人屬性變數(性別、年齡、職業)、住宅屬性(擔保品座落縣市、屋齡)及貸款條件屬性(房屋貸款金額、授信用途、貸款契約利率、貸款年限)等做為解釋提前償還的特徵變數，而是否提前償還則做為被解釋變數或輸出變數。各相關變數定義如表二，其中性別、年齡、職業、擔保品座落縣市、授信用途用虛擬變數D1~D5表示，而屋齡、貸款金額(取自然對數)、貸款時的契約利率、貸款年限則應用實際數值變數X1~X4表示。

表三列出各實質變數X1~X4之基本統計量，包含平均數及標準差，而表四則列出虛擬變數D~D5的次數分配表。從表三及表四中可以瞭解上述資料中所隱含的一些相關訊息：

1. 貸款人屬性方面：在性別中，男性佔48.2%，女性佔51.8%，二者相當接近。年齡方面，大部分是介於26-45歲(約佔69%)，表示購屋者在稍有經濟基礎後方進行購屋行為，符合經濟理性的抉擇。而貸款人職業則大部分從事工業、農林漁牧業及金融業、服務業(佔71%)，公教最少(僅佔3.8%)。
2. 住宅屬性方面：購屋貸款抵押擔保品座落在北部地區較多(佔72.4%)，顯示向該行貸款的購屋者之區位選擇集中於都市地區。屋齡平均約為1.34年，表示購屋者以偏好新成屋居多。
3. 貸款條件屬性方面：貸款金額大部分介於100萬至500萬間，取自然對數後平均數為6.36，授信用途呈現大部分的購屋者是以政府優惠房貸方式申貸(佔35.4%)及理財、指數型房貸(佔61.4%)。貸款契約利率平均為7.68%，貸款年限平均為19.66年。

(二) 實驗設計與結果分析

至於訓練樣本與測試樣本數目的選擇，不同的配對可能會產生不同的分類績效，但文獻上並未出現嚴謹的理論說明。本文按正常償還戶與提前償還戶以2:1隨機抽取的方式選取(註13)，在訓練樣本中，正常償還戶有1,200筆，提前償還戶有600筆，而在測試樣本中，則分別為600筆及300筆，詳細情形如下表五所示，

而在GP-DT的實證，本研究採用Discipulus™軟體，該軟體不但分類能力很強，且在參數設定上有較大的彈性。本文設定GP-DT的母體大小為500，表示每一個族群有500個個體(決策樹)，這大小對一般的分類問題已能達到不錯的績效(註14)，而函式集包含了+，-，×，%，sin，cos，EXP(指數)，終端集則包含了9個特徵變數及常數(註15)，交配率及突變率在此軟體中，可設為隨機調整模式，切割值分別設為0.3、0.5及0.8，其中0.5是文獻上最常設的值，而0.3及0.8則是做為比較之用；適合度函數則為誤差平方和(sum of squared error, SSE)，演化代數分別為50、250及500，目的亦是比較不同演化代數的績效是否存在顯著差異，詳細設定如表六所示。

Logit迴歸模型與一般線性迴歸之主要差異在於變數不若線性迴歸為連續性變數，而是以二分類變數形態出現，因此必須透過Logistic函數的轉換使其與預測變數呈現線性的關係，而

表二 變數說明

變數名稱及代號		變數說明	
貸款人 屬性	性別(D1)	男性：D1=1	女性：D1=0
	年齡(D2)	貸款人貸款時之年齡。	
		D2=1 (25歲或以下)	D2=2 (26至35歲)
		D2=3 (36至45歲)	D2=4 (46至60歲)
D2=5 (61歲以上)			
職業(D3)	D3=1 (職業為公家機關、教育機關)		
	D3=2 (職業為工業、農林漁牧業)		
	D3=3 (職業為金融業、服務業)		
	D3=4 (無或其他)		
住宅屬性 (D4)	D4=1 (北部地區)：台北縣市、桃園縣市、新竹縣市		
	D4=2 (中部地區)：台中縣市、彰化縣、雲林縣		
	D4=3 (南部地區)：嘉義市、高雄縣市、台南縣市		
屋齡(X1)	X1：貸款人申辦貸款時之屋齡		
貸款金額(X2)	X2：取自然對數		
貸款條件 屬性	授信用途(D5)	D5=1 (一般購屋)	D5=2 (房屋修繕)
		D5=3 (首次購屋)	D5=4 (整批購屋)
		D5=5 (政府優惠房貸)	D5=6 (員工購屋)
		D5=7 (理財、指數型房貸)	D5=8 (平轉房貸)
		D5=9 (專案房貸)	D5=10 (其他)
貸款契約利率(X3)	X3：貸款人申辦貸款時之契約利率		
貸款年限(X4)	X4：貸款人申辦貸款時之契約貸款年限		

表三 各實質變數之基本統計量

變數名稱	平均數	標準差
屋齡(X1) (年)	1.34	1.59
貸款金額(X2) (取log值)	6.36	0.34
貸款契約利率(X3) (%)	7.68	1.07
貸款年限(X4) (年)	19.66	1.68

表四 各虛擬變數之次數分配表

變數名稱	類別	次數	百分比	變數名稱	類別	次數	百分比	
性別(D1)	男：1	1398	51.8%	擔保品 座落縣市(D4)	1：北部地區	1956	72.4	
	女：0	1302	48.2%		2：中部地區	357	13.2	
年齡(D2)	1：(25歲以下)	101	3.7%		3：南部地區	387	14.4	
	2：(26至35歲)	936	34.7%		1：(一般購屋)	6	0.2	
	3：(36至45歲)	935	34.6%		2：(房屋修繕)	4	0.1	
	4：(46至60歲)	688	25.5%		3：(首次購屋)	6	0.2	
	5：(61歲以上)	40	1.5%		4：(整批購屋)	32	1.2	
職業(D3)	1：(公家、教育)	102	3.8		授信用途(D5)	5：(政府優惠房貸)	955	35.4
	2：(工、農林漁牧)	1920	71.1			6：(員工購屋)	1	0
	3：(金融、服務業)	396	14.7			7：(理財、指數型房貸)	1655	61.4
	4：(無或其他)	282	10.4	8：(平轉房貸)		9	0.3	
				9：(專案房貸)		32	1.2	
				10：(其他)		0	0	

表五 實證樣本設計

樣本	訓練樣本		測試樣本	
	正常償還戶	提前償還戶	正常償還戶	提前償還戶
類別				
數目	1200	600	600	300
百分比(%)	66.7	33.3	66.7	33.3
總計	1800		900	

表六 遺傳規畫決策樹參數設定

母體大小	500
函式集	{+, -, ×, %, sin, cos, EXP}
終端集	{D1, D2, D3, D4, D5, X1, X2, X3, X4, R}
交配機率	random
突變機率	random
臨界值	0.3、0.5、0.8
適合度函數	sum of squared error (SSE)
演化代數	50、250、500

在二分類的Logit模型中，經由Logistic函數的轉換可得出各變數與提前償還率 P_i 的關係。由 P_i 進而建立提前償還分類法則，若 P_i 大於切割值0.3 (或0.5、0.8)則歸類為1 (提前償還戶)，否則為0 (正常還款戶)，其關係如(6)式，

$$P_i = \begin{cases} 1 & \text{if } P_i \geq 0.3, 0.5, 0.8 \\ 0 & \text{otherwise} \end{cases} \dots\dots\dots (6)$$

為了瞭解不同切割值與GP-DT演化代數對模型評估績效的影響，底下做更進一步的分析。

1. 不同切割值的績效評估

大部分的文獻在評估分類模型的預測能力時，通常藉由模型的個別與整體正確率或誤判率高低衡量預測的效果，但這種二分類問題對切割值的選擇相當敏感，不同的切割值確實會影響分類的正確率，雖然大部分文獻以0.5為切割值，但如何建立嚴謹的理論基礎來證明何種切割值對分類的正確率仍有待努力，所以在實證設計上考慮了0.3、0.5、0.8三種不同的切割值，並進一步來比較其差異性。結果顯示出在Logit模型方面，無論樣本內或樣本外的整體平均正確率皆以0.5的切割值最佳、0.3次之、0.8時最差，這結果亦具有一致性。在GP-DT模型的整體平均正確率大小亦呈現一致性的現象，即無論切割值為何，均是GP-DT500優於GP-DT250，而GP-DT50的績效最差。而就不同種切割值間的比較，GP-DT的整體平均正確率是當切割值為0.5時最佳、0.3次之、0.8最差。

由上述的分析得知，不同切割值對分類模型的結果及正確率的評估，確實有不同的影響，但如何篩選切割值仍有待更多的實證驗證及更嚴謹的理論基礎。然而為克服切割值的影響，本文以ROC比率值的大小來比較模型在不同切割值下之分類結果，並評估分類模型的區別效果(註16)。

2. 不同演化代數的績效評估

理論上，演化代數愈多，愈能夠找出最佳的決策樹，亦表示有更好的分類績效，由本實驗結果表七至表九可以看出，就同一切割值而言，其分類績效(整體平均正確率)均以500代最佳，再來是250代及50代。而就不同種切割值間的比較，GP-DT的整體平均正確率是當切割值為0.5時達92.20%最佳，0.3次之，0.8最差，而GP-DT另一個值得注意的地方是在測試樣本的績效又比訓練樣本為佳，這是其他模型(如Logit)少有的現象。

圖四以整體平均正確率最高的GP-DT 500-3 (切割值為0.5，演化代數500代)說明收斂情形。由圖中可以看出，當演化代數逐漸增加時，無論是最佳正確率或是最佳5次的平均正確率均呈顯著增加，一直到第90代左右則出現收斂現象，整體正確率約92%左右，且無過度學習(overlearning)現象。而圖五則描畫出Logit及GP-DT500-3的ROC曲線圖(註17)。

因型一誤差所造成的成本損失比型二誤差更高，也就是將提前償還戶誤判為正常還款戶所造成的成本較將正常戶誤判為提前還款戶還嚴重，而本研究以測試樣本來進行效度指標比較，若由表十各模型的型一及型二誤差觀察，可以發現二者在大部分的模型中並無明顯的抵換關係(trade-off)，但GP-DT的型一及型二誤差均比Logit低。

3. ROC比率值及誤判成本分析

而就ROC比率值而言，由表十中可以看出，無論何種切割值或演化代數，不同GP-DT模

表七 切割值為0.3的績效評估

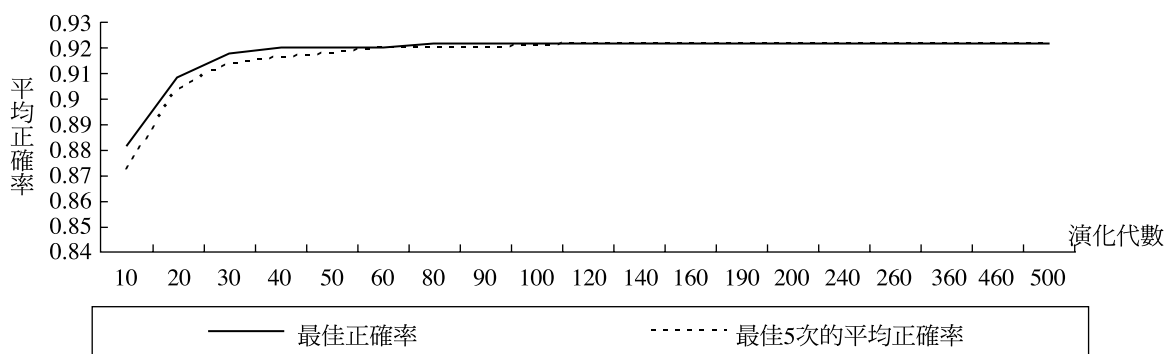
訓練樣本				
項目	Logit-1	GP-DT 50-1	GP-DT 250-1	GP-DT 500-1
整體正確率(%)	80.61	90.78	90.83	91.83
類別1正確率(%)	88.00	99.33	99.33	99.33
類別0正確率(%)	76.92	86.58	88.67	88.83
測試樣本				
項目	Logit-1	GP-DT 50-1	GP-DT 250-1	GP-DT 500-1
整體正確率(%)	76.67	91.28	92.22	92.33
類別1正確率(%)	76.33	99.33	99.33	99.33
類別0正確率(%)	76.83	88.00	88.67	88.83
整體評估				
整體平均正確率	78.64	91.03	91.53	92.08
績效排名	4	3	2	1

表八 切割值為0.5的績效評估

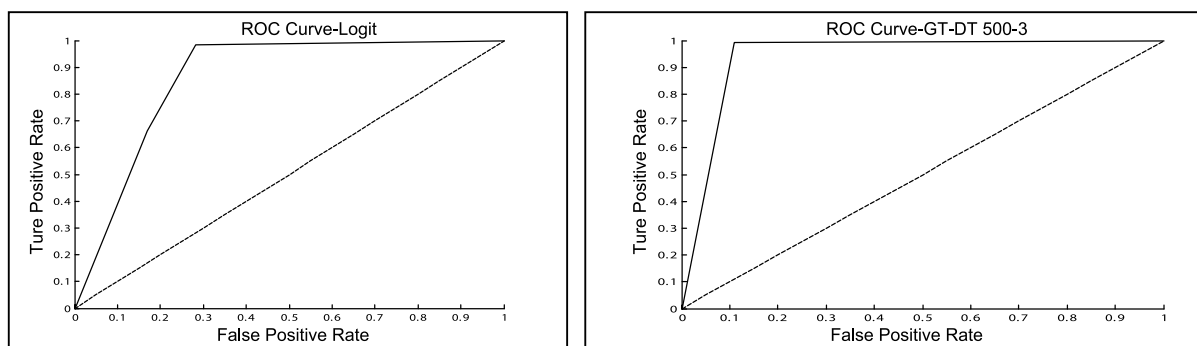
訓練樣本				
項目	Logit-2	GP-DT 50-2	GP-DT 250-2	GP-DT 500-2
整體正確率(%)	82.11	90.78	91.33	92.06
類別1正確率(%)	84.83	99.33	99.00	100
類別0正確率(%)	80.75	86.58	87.42	88.50
測試樣本				
項目	Logit-2	GP-DT 50-2	GP-DT 250-2	GP-DT 500-2
整體正確率(%)	75.78	91.78	91.89	92.33
類別1正確率(%)	65.33	99.00	99.33	99.33
類別0正確率(%)	81.00	86.58	87.50	88.83
整體評估				
整體平均正確率	78.95	91.28	91.61	92.20
績效排名	4	3	2	1

表九 切割值為0.8的績效評估

訓練樣本				
項目	Logit-3	GP-DT 50-3	GP-DT 250-3	GP-DT 500-3
整體正確率(%)	68.56	89.67	90.78	90.94
類別1正確率(%)	8.33	97.67	99.00	99.67
類別0正確率(%)	98.67	85.67	88.83	88.50
測試樣本				
項目	Logit-3	GP-DT 50-3	GP-DT 250-3	GP-DT 500-3
整體正確率(%)	66.22	91.00	91.67	91.78
類別1正確率(%)	2.00	98.33	99.33	100.00
類別0正確率(%)	98.33	87.33	87.50	88.00
整體評估				
整體平均正確率	67.39	90.34	91.23	91.36
績效排名	4	3	2	1



圖四 GP-DT-500-3收斂圖



圖五 Logit與GP-DT 500-3的ROC曲線

表十 各模型的ROC比率值及誤判成本

模型簡稱	ROC 比率值	型一 誤差	型二 誤差	誤判成本 (成本率=0.5)	誤判成本 (成本率=1.0)	誤判成本 (成本率=1.5)	誤判成本 (成本率=2)
Logit-1	0.8389	23.67	23.17	23.27	23.34	23.38	23.42
GP-DT 50-1	0.9134	0.67	12	9.73	8.22	7.14	6.34
GP-DT 250-1	0.9203	0.67	11.33	9.20	7.78	6.76	6.00
GP-DT 500-1	0.9345	0.67	11.17	9.07	7.67	6.67	5.92
Logit-2	0.8389	34.67	19	22.13	24.22	25.72	26.84
GP-DT 50-2	0.9165	1	13.42	10.94	9.28	8.10	7.21
GP-DT 250-2	0.9281	0.67	12.5	10.13	8.56	7.43	6.59
GP-DT 500-2	0.9385	0.67	11.17	9.07	7.67	6.67	5.92
Logit-3	0.8389	98	1.67	20.94	33.78	42.95	49.84
GP-DT 50-3	0.9202	1.67	12.67	10.47	9.00	7.96	7.17
GP-DT 250-3	0.9245	0.67	12.5	10.13	8.56	7.43	6.59
GP-DT 500-3	0.9464	0	12	9.60	8.00	6.86	6.00

說明：GP-DT50-1代表演化代數為50代，切割值為0.3；GP-DT50-2代表演化代數為50代，切割值為0.5；GP-DT50-3代表演化代數為50代，切割值為0.8；Logit-1代表切割值為0.3；其餘依此類推。

型均一致性地優於Logit模型，當考慮成本率的變化而產生的誤判成本損失時，若單以整體誤判率或以ROC比率值的比較並無法正確地評估模型的好壞。因此，本研究進一步以測試樣本的誤判成本來衡量各模型的優劣性，然就誤判成本來看，本文根據所建立的模型，以模擬的成本率來探討其變化所造成的影響，分別設定了0.5、1、1.5、2等四種不同的成本率，誤判成本愈高，表示模型預測的結果所造成的損失幅度愈大，由表中的誤判成本可以明顯看出，不同GP-DT模型仍一致性地優於Logit模型，且除了Logit外，GP-DT各模型的成本率增加時，其誤判成本均一致性地降低，而以GP-DT 500-2 (成本率為2)最低，理由是本研究的提前償還樣本僅是正常還款戶的一半，而且各模型的型一誤差大部分低於型二誤差，以致當成本率增加時，誤判成本公式的分子項增加的數值小於分母項，因而造成誤判成本降低的現象。

4. GP-DT決策法則的形成

我們進一步討論GP-DT決策法則的形成，因Discipulus™軟體能自動演化提前償還的預測模型，並可將程式轉換成易懂的C語言程式，接著將每位貸款戶的特徵變數以不同的法則來判斷，最後並計值提前償還機率值 P_i ，若 P_i 值大或等於切割值(0.3, 0.5, 0.8)，則歸類為1 (提前償還戶)，反之，則為0 (正常還款戶)。

5. 特徵變數的敏感度分析

為瞭解各特徵變數對提前償還的影響程度，本研究參考李沃牆、陳義先(2004)的作法及廖慶榮(2006)的建議，以敏感度分析(sensitivity analysis)來進行評估。敏感度分析係指當刪除

某一輸入變數時所導致模型測試結果增加之誤差百分比或減少正確率之百分比，以此來說明該輸入變數的相對重要性。本研究應用全樣本資料，並以切割值為0.5的Logit模型及績效較佳的GP-DT 500-3模型進行比較，結果如表十一所示。以其中的性別變數而言，當Logit模型忽略了此一變數，即只考慮其他8個變數，則其正確率為81.72%，同理，GP-DT 500-3模型為90.23%，將此二模型加權平均，如Logit模型權重為0.4752，而GP-DT模型為0.5248，則可獲得平均加權正確率為86.1855%，而表中最後一欄則是此變數的重要性排名，平均加權正確率愈高，表示此一變數的重要性愈低，所以由表中可看出，這些特徵變數在此二模型中的影響結果相當接近，而經加權後，重要性排名順序依次為授信用途、擔保品座落縣市、職業、貸款契約利率、性別、年齡、貸款金額、屋齡、貸款年限。

本文之敏感度分析或許在方法上嚴謹度較低，但對於瞭解個別變數對提前清償的影響程度時，仍不失為一種可行的分析方法。

由本文的分析，可據以建構房貸的風險管理系統，如圖六所示。首先是將影響提前償還的因子輸入GP-DT提前償還預警系統，如果預測結果會提前償還，則可估算因提前償還而損失的現金流量，而欲先規避此類風險，則可透過提前償還違約金的設計來避險(註18)；另一方面，若預測結果不會提前償還，則應考慮是否為正常繳款或是逾期繳款，對此部分則有賴於房貸審核系統的建構。

六、結論

住屋貸款中最主要的風險為違約與提前償還風險，因為這兩個原因均會造成貸款支付的停止或是現金流量的不確定。如何規避這些風險、降低不確定性與損失，並建構一套嚴謹且有效的房貸風險管理系統是有其必要的。

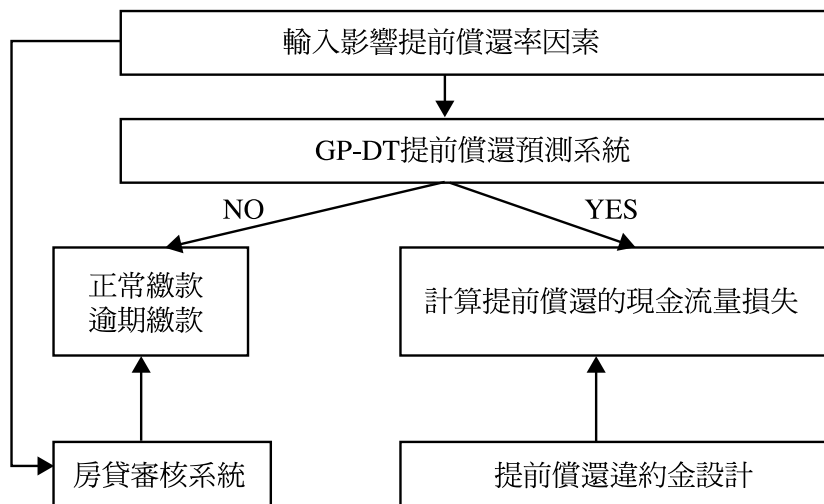
針對上述的議題，本文利用無母數的遺傳規畫決策樹模型來預測提前償還率，並與Logit模型比較，進而建構房貸風險管理系統。實證結果得出，無論切割值及演化代數為何，無母數的GP-DT之預測正確率均優於Logit模型，理由是無母數的模型不需要太多嚴謹的假設，並且具有自我學習及演化的功能，尤其是對提前償還率決策法則形成有明確的解釋，所以在此議題上的應用相當適合。而考慮不同的臨界值，Logit模型無論在樣本內外均無一致性，但以0.8的切割值最差，而就GP-DT模型而言，則出現一致性的結果，即無論切割值為何，GP-DT500的整體平均正確率均最佳，若進一步比較不同切割值的個別績效差異，則以0.5的切割值最好。而由不同的演化代數結果顯示，愈高的演化代數能夠獲得充裕的學習，所以在本文的GP-DT模型中，只要不出現過度學習，500代的演化次數均能呈現較佳的結果。

本研究亦根據所建立的模型，以模擬的成本率來探討其變化所造成的影響，分別設定了0.5、1、1.5、2等四種不同的成本率，結果得出，不同GP-DT模型仍一致性地優於Logit模型，且除了Logit模型外，GP-DT各模型的成本率增加時，其誤判成本呈現一致性地降低，而以GP-DT500-1、GP-DT500-2且成本率為2時最佳。

然瞭解各特徵變數對提前償還率影響程度對於房貸的風險控管有絕對的幫助，本文利用敏感度分析，並透過Logit模型及GP-DT模型的整體正確率來決定權重，以解釋各特徵變數的重要性，這結果無論對房貸的審核或提前償還的預測均有一定的參考價值。最後，本文提出

表十一 特徵變數的敏感度分析

變數類別	刪除變數名稱	Logit模型 正確率(%)	GP-DT模型 正確率(%)	平均加權 正確率(%)	重要性 排名
貸款人屬性	性別(D1)	81.72 (0.4752)	90.23 (0.5248)	86.1855	5
	年齡(D2)	81.83 (0.4745)	90.63 (0.5255)	86.4545	6
	職業(D3)	77.83 (0.4767)	85.42 (0.5233)	81.8014	3
住宅屬性	擔保品座落縣市(D4)	75.83 (0.4794)	82.33 (0.5206)	79.2135	2
	屋齡(X1)	82.23 (0.4762)	90.46 (0.5238)	86.5411	8
	貸款金額(X2)	82.11 (0.4756)	90.52 (0.5244)	86.5198	7
貸款條件屬性	授信用途(D5)	72.33 (0.4928)	74.45 (0.5072)	73.4053	1
	貸款契約利率(X3)	82.17 (0.4864)	86.76 (0.5136)	84.5273	4
	貸款年限(X4)	82.28 (0.4760)	90.57 (0.5240)	86.6238	9



圖六 房貸授信決策風險管理系統

房貸風險管理系統，除了可降低房貸的風險及損失外，更可做為銀行進行資產負債管理的一個子系統，尤其Basel II已於2007年開始實施，銀行更應及早做準備。

信用資料的取得原就相當不易，而本文的實證資料受限於提供單位所能擷取的變數相當有限，影響提前償還的諸多特徵變數無法取得並納入本文的研究，對實證結果可能會有影響，再來是因住宅貸款戶還本付息的時點為按月償還，銀行未來若能建立完整的還款資料庫，則將來的研究變數選取就能以月資料做探討，或許更能捕捉借款者的動態還款行為，更能促進模型的預測能力。另一個議題，則是針對提前償還的違約金設定進行實證研究。而在實證方法上，後續研究亦可考慮Cox(1972)的存活分析模型(survival analysis model)來探討不同時點的提前償還行為；而不同的提前償還與正常還款戶分類比例是否對結果產生差異，亦是

後續值得研究之處；在績效的測度上，亦可考慮交叉驗證(cross-validation)方式，最後，針對總體變數變動及首次購屋之優惠利率(teaser rate)是否會對提前償還產生影響亦可加入後續研究討論。

註 釋

- 註1. 消費者貸款業務包括購置住宅貸款、房屋修繕貸款、購置汽車貸款、機關團體職工福利貸款，以及不含信用卡循環信用之其他個人消費貸款。
- 註2. 資料顯示，自2000年以來，央行總計推出六次優惠房貸，每次可貸款額度在2000億元至3000億元，累計貸款總額已超過1.6兆元台幣。
- 註3. Abrahams(1997)指出，住宅貸款的提前償還包括：重新融資(refinance)、換屋(house turnover)、違約(default)、及部份提前償還(curtailment)。
- 註4. CIR模型是Cox et al.(1985)所提出的模型。
- 註5. due-on-sale條款是銀行與住宅貸款債務人間的債務條款，規定當債務人出售房屋時，銀行有權要求債務人立即將剩餘的住宅貸款還清。
- 註6. 非線性的分類模型尚包含類神經網路(artificial neural networks, ANN)、模糊邏輯(fuzzy logic)等，但這些模型不是對決策的形成不易解釋，就是歸屬函數的設定不甚明確。
- 註7. 決策模型係指繼續付款、提前清償或者違約。
- 註8. 詳細的遺傳規畫介紹，有興趣的讀者可參考Koza(1992)的著作。
- 註9. 至於選擇及突變運算，文獻上有很多詳述，本文不再贅述，有興趣的讀者亦可參考Koza(1992)。
- 註10. CAP曲線可透過正確率(AR)來衡量其模型的區別力，且AR值會介於0~100%之間，愈接近100%表示該模型之預測能力愈佳。
- 註11. ROC曲線即圖形上的彎曲曲線(如圖三所示)，而面積A即ROC曲線與45度線所夾的面積(亦即ROC曲線下方區域)，稱為ROC比率(簡稱AUROC)。當ROC比率等於0.5時，即表示該模型為無任何區別力的隨機模型；而比率為1時4，則表示該模型為完美模型。故愈接近1表示該模型之預測能力愈佳。
- 註12. 成本率係指因誤判產生之型一誤差與型二誤差所導致的損失成本之比率。舉例來說，成本率為10即表示因型一誤差所造成的損失成本為型二誤差的10倍。由(5)式及上述成本率的定義，我們可以發現，成本率分別在A及C的定義中出現，若型一誤差大於型二誤差，則型一誤差乘上成本率的結果會使誤判成本增加。
- 註13. 實證上，提前或正常還款戶應以多少比例，文獻上並無定論，大部分文獻設某一特定比例的目的均為提高模型的判別績效；如Altman(1968)將正常和違約公司以1：1的比例進行實證，而國內的沈中華、忻惟毅(2006)將正常和違約公司以4：1的比例進行實證。而本文所用的2：1，理由是提前清償戶樣本原本較少，而為提高模型學習及預測能力，因而採取此比例。
- 註14. 遺傳規畫(GP)是由Koza(1992)所提出，在該書中，Koza對每一代族群的設定大小(population size)大多設為500，對各種相關問題的解決(包括分類問題，投資策略，一般的多項式問題……)均可達到不錯的績效，因為此值太小則不易收斂，太大則容易提早

收斂，而易掉入局部最適解，其中如Devaney & Hagedorn(2002)亦是設為500。

註15. 常數值介於[-1,1]之間。

註16. 切割值以何者為最適，文獻上並無一致性的結論，但都希望所取的臨界值能提升模型判斷的正確率，如紀麗秋(2002)的研究將危機發生機率以0.1至0.9，每間隔0.1為臨界點(cutoff point)，以降低型一誤差為考量，找出區別正確率最高者為最適臨界點，因此該研究建構之模式在各年度的臨界點均不同。

註17. 因GP-DT各模型的ROC曲線非常接近，本研究僅列出以整體誤判率較低的GP-DT500-3模型。

註18. 有關房貸提前償還違約金的設計，李沃牆、彭敏瑜(2007)有進一步的討論。

參考文獻

沈中華、忻惟毅

2006 〈中小企業違約機率的預測〉《金融風險管理季刊》2(1)：97-114。

李沃牆、彭敏瑜

2007 〈房貸提前償還率違約金之探討〉《永豐金融季刊》39：93-120。

李沃牆、陳義先

2004 〈信用卡審核評分表之建構〉《真理財經學報》10：23-56。

何棟欽

2008 〈美國次級房貸危機：證券化風險之啟示〉《永豐金融季刊》41：55-112。

施恩

1994 《住屋貸款還款速度之研究》碩士論文，國立台灣大學商學系。

紀麗秋

2002 〈出口貿易信用風險模型之研究—以亞太地區為例〉《台灣金融財務季刊》3(1)：81-116。

郭姿伶

2000 《住宅貸款之提前清償與逾期還款》碩士論文，國立中正大學財務金融研究所。

陸文傑

2001 《抵押貸款證券之評價—Implied Prepayment 之應用》碩士論文，國立台灣大學財務金融研究所。

陳建勝、林明宏

2003 〈壽險保單早期失效之預測〉《風險管理學報》5(3)：341-361。

陳健豪

2003 《台灣地區住宅抵押貸款—提前清償與違約之分析》碩士論文，國立清華大學科技管理研究所。

楊顯爵、王昭文、劉君德

2006 〈房屋修繕貸款群組的提前清償率比較—指數迴歸法〉《臺灣金融財務季刊》7(3)：1-33。

黃文啟

2002 《以 LOGIT 模型研究借款人特性與不動產抵押貸款提前償還之關係》碩士論文，國立政治大學財務管理學系。

黃朝基

2003 《不動產抵押貸款提前償還影響因素之研究》碩士論文，私立逢甲大學土地管理研究所。

黃建智

2004 《以卜瓦松迴歸方法探討房屋抵押貸款提前償還與違約決策》碩士論文，國立政治大學企業管理學系。

程博平

2003 《影響企業提前償還率因素之探討》碩士論文，私立文化大學會計系。

劉展宏、張金鶚

2001 〈購屋貸款提前清償行為之研究〉《住宅學報》10(1)：29-49。

廖慶榮

2006 《作業研究》台北：華泰書局，修訂一版。

簡士龍

2003 《住宅抵押貸款事務性服務權之評價》碩士論文，國立清華大學科技管理研究所。

Abrahams, S. W.

1997 “The New View in Mortgage Prepayments: Insight from Analysis at the Loan-by-Loan Level,” *The Journal of Fixed Income*. 7(1): 8-21.

Ambrose, B. W. & M. LaCour-Little

2001 “Prepayment Risk in Adjustable Rate Mortgages Subject to Initial Year Discounts: Some New Evidence,” *Real Estate Economics*. 29(2): 305-327.

Ambrose, B. W. & A. B. Sanders

2001 “Commercial Mortgage-Backed Securities: Prepayment and Default,” Working Paper.

Archer, W. R., D. C. Ling & G. McGill.

1997 “Demographic Versus Option-Drive Mortgage Terminations,” *Journal of Housing Economics*. 4: 1051-377.

Boonyanunta, N. & P. Zeephongsekul

2000 “State of the Art Credit Risk Analysis Model: Comparative Analysis Between Statistical Approaches and Neural Network Approaches,” Unpublished paper presented in the Fifth Conference of the Association of Asian-Pacific Operations Research Societies within IFORS, Singapore.

Breiman, L., J. H. Friedman, R. A. Olshen & C. J. Stone

1984 *Classification and Regression Trees*. New York: Chapman & Hall.

Campbell, T. S. & J. K. Dietrich

1983 “The Determinants of Default on Insured Conventional Residential Mortgage Loans,” *Journal of Finance*. 38(5): 1569-1581.

Clapp, J. M., Y. H. Deng & X. D. An

2006 “Unobserved Heterogeneity in Models of Competing Mortgage Termination Risks,” *Real Estate Economics*. 34(2): 243-273.

Cox, D. R.

1972 “Regression Models and Life-Tables,” *Journal of the Royal Statistical Society, Series B*. 34(2): 187-220.

Cox, C., J. E. Ingersoll & S. A. Ross

1985 “A Theory of the Term Structure of Interest Rates,” *Econometrica*. 53: 385-407.

Cunningham, D. F. & C. A. Capone

1990 “The Relative Termination Experience of Adjustable to Fixed Mortgages,” *Journal of Finance*. 5(5): 1687-1703.

- Devaney, J. E. & J. G. Hagedorn
2002 "The Role of Genetic Programming in Describing the Microscopic Structure of Hydrating Plaster," Genetic and Evolutionary Computation Conference (GECCO-2002).
- Dunn, K. B. & J. J. McConnell
1981 "A Comparison of Alternative Model for Pricing GNMA Mortgage-Backed Securities," *The Journal of Finance*. 36(2): 471-487.
- Engelmann, B., E. Hayden & D. Tasche
2002 "Measuring the Discriminative Power of Rating Systems," Working Paper, Deutsche Bundesbank, University of Vienna.
- Green, J. R. & J. B. Shoven
1986 "The Effects of Interest Rates on Mortgage Prepayments," *The Journal of Money, Credit, and Banking*. 1(1): 41-59.
- Kau, J. B., C. Keenan, W. J. Muller & J. F. Epperson
1990 "The Valuation and Analysis of Adjustable Rate Mortgages," *Management Science*. 36(12): 1417-1431.
- Koza, J. R.
1992 "Genetic Programming: On the Programming of Computers by Means of Natural Selection," Cambridge: MIT Press.
- LaCour-Little, M.
1990 "Another Look at the Role of Borrower Characteristics in Predicting Mortgage Prepayments," *Journal of Housing Research*. 10: 45-60.
- Quigley, J. M.
1987 "Interest Rate Variations, Mortgage Prepayments and Household Mobility," *The Review of Economics and Statistics*. 69(4): 636-643.
- Quinlan, J. R.
1983 "Learning Efficient Classification Procedures and Their Application to Chess end Games," in *Machine Learning: An Artificial Intelligence Approach*. 463-482. ed. R. S. Michalski, J. G. Carbonell, & T. M. Mitchell. Palo Alto: Tioga.
- Smith, L. D., S. M. Sanchez & E. C. Lawrence
1996 "A Comprehensive Model for Managing Credit Risk on Home Mortgage Portfolios," *Decision Sciences*. 27(2): 291-317.
- Van Deventer, Donald R. & Xiaoming Wang
2003 "Measuring Predictive Capability of Credit Models Under the Basel Capital Accords: Consecro and Results from the United States, 1963-1998." Kamakura Working Papers (<http://www.kamakuraco.com>), WP16.

