澳 門 科 技 大 學
MACAU UNIVERSITY OF SCIENCE AND TECHNOLOGY

學士學位作業報告

基於 Boosting 提升算法的房屋租賃意願預測

學生編號	學生姓名
1709853V-B011-0053	谷奕霖

學　　院：商學院

課程名稱：工商管理學士學位課程

指導老師：趙鴻昊

遞交日期：2021 年 3 月 12 日

摘要

數據二字在 21 世紀成爲出現頻率最高的字眼，無論在哪個產業，數據所扮演的角色逐漸變得越發重要，而它的重要性不僅僅在其本身，更在於科學分析及建模后為決策者帶來的啓示。

本文將著眼與美國紐約的房地產行業，通過特徵工程和建模，對租房者的租房意願進行預測。我們將要處理的是一個三分類問題，租房意願被分爲低中高三類。模型的自變量包括並不局限與：經緯度，房型特徵(如房間個數等)，價格，信息錄入時間等。由於數據集分佈不均衡（意願高的一類占比僅有 7.78%），通過不同采樣方法的比較，我們最終選用了 SMOTEENN 法進行過采樣。在特徵篩選方面，通過對比過濾法，包裹法與嵌套法三大類特徵篩選方法，最終選擇了嵌套法中的随机森林算法。在模型選擇方面，基於 Boosting 算法我們分別構造了 AdaBoost, GBDT 與 XGBoost。在不斷的嘗試當中，我們發現随机森林与 GBDT 算法的組合对于该问题的表現最佳，然而，在三分类问题的背景下，采样法与 Boosting 算法的结合反而对结果产生了干扰，降低了预测模型的准确率，这一点笔者将其归为 Boosting 算法的特点之一。

关键词：分類预测；機器學習；房屋租賃；Boosting 算法；SMOTEENN

Abstract

Data has become the most frequently used word in the 21st century. No matter in which industry, the role of data has gradually become more and more important, and its significance lies not only in data itself, but in the value and business insights revealed from the entire decision-making process through scientific analytics and modeling.

This article will focus on the real estate industry in New York, USA, and predict the renters' willingness through feature engineering and modeling. What we are going to deal with is a multi-classification problem. For the object variable, the willingness to rent a house is divided into three categories: low, medium, and high, while independent variables of the model include geographical features such as latitude and longitude, room characteristics (such as the number of rooms, etc.), price, etc. Due to the uneven distribution of the dataset (only 7.78% of the willing ones), we finally chose the SMOTEENN method for oversampling through the comparison of different sampling methods. In terms of feature engineering, Random Forest in Embedding approach was selected by comparing three types of methods: filtering method, wrapping method and embedding method. For model selection, we constructed AdaBoost, GBDT and XGBoost respectively based on the Boosting algorithm. In continuous attempts, we found that the combination of Random Forest and GBDT performs best. However, in the context of the three-classification problem, the combination of sampling method and Boosting algorithm interferes with the results and reduces the accuracy of the classification model. The author attributed this phenomenon to one of the characteristics of the Boosting algorithm and modified the evaluation function to better present the results.

Key words: Classification Models; Machine Learning; House Renting; Boosting Algorithm; SMOTEENN Sampling

目录

第一章 前言

1.1 社會背景

從古至今，房子在人們的心中占据着不可替代的地位，對於每個家庭都有著十分重要的意義。現代社會中，科學技術迅猛發展，人們的住房在性能、功能上有了新的突破。與此同時，人口增長迅速，流動性加大，房屋的需求也大大增加。然而，在房地產成為中國最重要的支柱產業時，一些問題也隨之而來。其中最大的問題是高昂的房價。尤其是在一些大城市，昂貴的房價成為許多人生活的一大負擔。根據賢集網的資料，2020 年全國房屋均價（元/平方米）的前十名為：1. 深圳：65516； 2. 北京：63052； 3. 上海：54467； 4. 廈門：46679； 5. 三亞：36533； 6. 廣州：35726； 7. 南京：31045； 8. 杭州：30543； 9. 天津：26325； 10. 福州：26021。以第 7 名南京為例，如果一個家庭想買一套 110 平方米的公寓，那麼需要約 341 萬人民幣。2020 年南京城鎮非私營單位從業人員年平均工資約為 12 萬人民幣。假設這個家庭年收入 24 萬，不做任何消費也需要 14 年多才能買下這套房屋。同時，城市化的快速推進，吸引了大批外來人口向佔據資源優勢的一、二線城市湧入。對於這些外來人口來說，在大城市買房顯得更加困難。因此，對於房屋租賃的需求開始不斷增加。

據統計，目前中國房屋租賃人數已超 2 億。在這其中，外來務工人員的租房需求尤為旺盛。同時，時而出現的房屋租賃市場亂象，讓租房的需求者面臨租房難的問題，這也成為了當下城市公共住房保障建設的一個痛點。再觀美國，自 08 年金融危機之後，美國租房人群比例上升。2018 年，美國租房率已達 39%。在租房人群中，30 歲以下的年輕租客在穩定增長，50-70 歲中老年增長迅速。由此可見，租房已經和許多公民息息相關。

然而，人們常常面臨租房難的困境。主要的問題是資訊不對稱，人們擔心資訊的真實性和完整性。隨著互聯網的發展，越來越多的租客會在房屋租賃平臺上挑選房屋，例如美國 RentHop 的線上房屋租賃網站、中國的 58 同城 APP 等等。然而，平臺上的房源眾多，品質參差不齊，還可能存在欺詐。如果能夠解決這些問題，那將使租房不再是一件難事，從而為廣大租房者帶來福音，提高居民的生活水準。

1.2 研究的動機與意義

本文研究的動機是解決人們租房難的問題。作為一個地理面積較小的現代化都市，澳門的租房問題一直引起社會的廣泛關注。借由澳門在讀生的身份，我們

切身感受到了租房問題的緊迫性，包括房屋資訊雜亂、資訊來源不明等問題。同時，租房者還將面臨欺詐等風險。因此，筆者希望通過專業所學，利用機器學習等知識理論，基於租賃房源資訊設計高效合理的模型來對租房意願進行預判，使公寓搜索更加智慧化，提升租房效率及安全性。同時，也能幫助租房平臺更早地發現房源潛在的品質問題，提高盈利。除了上述的現實意義，該研究也具有一定的理論意義，即：本次论文涉及三大特徵篩選的方法——過濾法、包裹法和嵌套法，并通过過採樣欠採樣處理样本，同时還嘗試多種 Boosting 算法，並對 Boosting 模型進行了改進和提升。

1.3 創新點

在對特徵處理後，我們的特徵數量達到了 1500 多個，所以特徵篩選顯得十分必要。為了提高我們對於房屋租賃意願預測的準確度，我們嘗試了過濾法、包裹法和嵌套法這三大類特徵篩選方法，通過實踐我們對比了各個方法的效果。我們的目的在於解決一個三分類問題，但是數據的不平衡屬性令我們的模型喪失了客觀性，因此我們在變量篩選之後對數據進行了過採樣欠採樣處理，這一點對於我們建模有非常重要的意義。基於 Boosting 算法，我們對處理后的數據進行了 AdaBoost, GBDT 與 XGBoost 的建模。由於我們最關注的是租房意愿高的数据，因此我们對目标函数進行調整，使模型更適合處理本文要解決的問題。

1.4 文章結構

本文分為七個章節，每個章節的內容安排如下：

第一章介紹了本文研究的社會背景，研究的動機與意義，該研究的一些創新點以及全文架構。

第二章是文獻綜述，對目前國內外部分關於租房意願的主流研究進行評述與總結。

第三章對文本的研究所使用的算法進行了較為詳細的介紹，包括特徵篩選的三大方法包括過濾法、包裹法和嵌套法，欠採樣過採樣和 Boosting 的相關算法。

第四章介紹數據來源以及對於數據的描述性分析。

第五章將描述數據的預處理過程以及特徵工程。

第六章介紹了整個建模過程，以全樣本預測正確率以及三種意願等級各自的預測正確率作爲評判標準，比較不同 Boosting 模型的表現，最後將不同特徵篩選

方法與模型的搭配結果以图表的方式進行對比和分析，根據相應的情景得出最優的算法搭配。

第七章是對全文的總結以及展望，並對本文的局限性進行了闡述。

第二章 文獻綜述

對比我國租房市場，西方租房市場起步較早，尤其是美國，發展相對更為迅速。美國從 20 世紀 70 年代就已經開始對租房市場進行研究，而我國在 20 世紀 80 年代以前，即改革開放前，是不在市場經濟體制之下的，因此我國租房市場的發展在此之前處於停滯狀態。隨著國內房地產業的逐步發展和完善，21 世紀初，國內對於租房市場的研究給予了更高的关注。綜合國內外租房消費市場的相關文獻，我們發現研究的主要論點集中在租房意願的影響因素、租房的供需關係與消費群體、政策對租房市場的影響等。其中，國內外對於租房意願的研究最為廣泛。基於本文的研究主題，我們將從租房意願的角度對國內外以往的研究成果進行梳理和綜述。

（1）國外租房意願的研究現狀

Kain 和 Quigley（1972）兩人開創了租房意願研究的先河。他們以家庭為單元，在租購房的選擇之間進行對比，對影響租房意願的因素進行分析，涉及的相關變數包括家庭規模、家庭組成、家庭成員就業情況等。研究表明：單身男性選擇租房的可能性大於單身女性；黑人租房的可能性大於白人；家庭規模越大，會更傾向于購房而非租房；收入水準和退休狀況與租房意願呈負相關 (Kain & Quigley, 1972)。

後來 Quigley（1976）首次將 Logit 方法用於租房意願研究，這是分類理論在租房意願研究中的首次運用，隨後很多類似的租房意願實證研究皆基於該方法進行 (Quigley, 1976)。Li（1977）提出有關租房自有率的 Logit 模型，研究波士頓地區居民的租房意願，研究變數涉及到戶主年齡、收入、家庭規模和種族等，依據 Logit 模型分析得出戶主年齡、家庭規模和收入水準與租房意願呈正相關，且黑人租房意願較白人更為強烈 (Li, 1977)。Silberman，Yochun 和 Inlanfeldt（1982）運用 Probit 模型研究了曼哈頓地區居民的租房意願，他們驗證了 Li（1977）的結論，即租房意願上白人家庭和黑人家庭是顯著不同的，黑人租房意願明顯更強烈，且年齡較大的黑人比較小的黑人租房意願更為強烈 (Silberman, Yochun & Inlanfeldt, 1982)。但 Shear, Watch 和 Weicher（1988）認為黑人和西班牙人租房意願並不如白人強烈，除此之外，他們還認為年齡較小、未婚、低收入及家庭規模小的家庭租房意願更強烈。由上得知，上世紀 80 年代諸多學者主要研究家庭特徵對租房意願的影響 (Shear, Watch & Weicher, 1988)。

進入 90 年代以後，租房意願的研究維度變得豐富起來。Jeffrey M.Perloff（1991）以人口結構為指標研究了外來務工人員的租房意願問題，結果表明年輕、已婚、文化程度低的人租賃意願較強 (Jeffrey M.Perloff, 1991)。Vander Hart（1994）專門對老年人的租房意願問題進行了研究，研究發現退休人員、殘疾人、孤寡老人傾向于租房，而已婚已育的老年人傾向于購房 (Vander Hart, 1994)。而 Eppli 和 Childs（1995）則研究了年輕群體的租房意願，研究表明婚姻因素與租房意願呈負相關，切嬰兒的出生會降低年輕人的租房意願 (Eppli & Childs, 1995)。Bourassa（1995）加入了房價租金比這一新變數，表明房價租金比、居民支出占居民收入比重增加會推動居民租房 (Bourassa, 1995)。Vander Hart（1998）和 Kan（2000）都使用了美國收入動態追蹤資料，Vander Hart 通過構建 Logit 模型分析得出戶主退休的家庭租房意願更強，以及戶主是否殘疾對租房意願沒有顯著影響 (Vander Hart, 1998; Kan, 2002)。

進入 21 世紀後，關於租房意願的研究維度進一步得以拓展，研究人員通過對租房家庭特徵、租賃成本、支付能力、心理因素、政策和制度等多方面因素進行多維度研究，對不同研究物件的租房意願研究業更為徹底全面。Boehm，Schlottmann（2004）對美國洛杉磯地區的租戶進行了研究，通過問卷走訪的方式，收集了租金、學歷、職業、收入、家庭人口、家庭搬遷次數等研究變數，研究發現受教育程度高、未婚的年輕群體租房意願最強；隨著租客年齡、收入的增加，租房意願逐步減弱 (Boehm & Schlottmann, 2004)。Beisky（2013）運用了 Logistic 模型，著重對租金、收入等財務類因素對於青年群體租賃意願的相關程度進行分析，結果表明即便房價上漲速度遠高於租金，青年群體的租房意願並不強烈。在社會住房供給不足的情況下，更多青年人傾向於在工作後和父母同住 (Beisky, 2013)。Thomas J（2017）以瑞典居民為研究对象，從經濟學的角度衡量了不動產和出租屋的實際市場價值，發現失業風險等不確定性因素會提升出租屋的實際市場價值，加大了房東的住宅持有意願，從而間接降低了瑞典居民的租房意願 (Thomas J, 2017)。

（2）國內租房意願的研究現狀

曾珍，邱道持，李鳳，李小廣（2012）和陳曉妍（2013）都運用 Logistic 模型定量分析了影響大學畢業生對租房意願的因素，並應用了隨機森林算法對相關影響因素進行貢獻度排序。曾珍, 邱道持, 李鳳, 李小廣（2012）認為未來定居城市、租房附近就業條件和當前職位及其工作地點對大學畢業生租房意願影響最為顯

著 (曾珍, 邱道持, 李鳳 & 李小廣， 2012)；但陳曉妍（2013）認為當前職業對租房意願無顯著影響，畢業年限、人均居住面積的影響更為顯著 (陳曉妍, 2013)。

基於上海較為成熟的租房市場環境，彭秀明（2013）總結了上海租賃住房的供給與需求狀況，並通過 Logit 模型分析出租房意願的影響因素主要為租金和住房條件，並用灰色 GM(1,1)模型對上海市外來人口租房意願進行預測 (彭秀明, 2013)。

路征，楊宇程，趙唯奇（2016）對 12 個城市的外來務工人員的租房特徵進行調查，選取 Probit 模型和 Ordered Probit 模型展開實證研究，研究表明：年齡、收入、申請門檻、租金、房屋結構、定居打算、情感需求等因素對租房需求有顯著影響，其中年齡、租金、申請標準、房屋結構的影響程度較大，收入水準更高的外來務工人員更傾向于通過購買住房來解決居住問題，需求意願相對更弱 (路征, 楊宇程 & 趙唯奇, 2016)。

莊靜，李夢微（2018）以南京市的應屆大學生為研究群体，通過抽樣調查、發放問卷的方式獲取資料，構建多元有序 Logistic 模型分析發現租房意願與對租房政策的瞭解程度、租房附加的公共服務呈顯著正相關關係，與租房的便利性、對租房政策的認可度呈顯著負相關關係 (莊靜 & 李夢微, 2018)。

孫曉輝，劉寶貞，王婷婷，劉璿（2019）從心理因素和社會制度因素出發，歸納了我國中低收入者租房意願逐年下降、貸款購房意願逐年上升的原因：現有社會制度使擁有房產權的人更容易獲得額外利益，例如社會保障、子女教育、公務員任職資格等；在住房價格不斷上漲時，決策者尤其是文化程度相對較低的中低收入者的認知常常出現系統性偏差；社會和心理綜合作用下導致的居民對安全的高度重視 (孫曉輝, 劉寶貞, 王婷婷 & 劉璿, 2019)。

劉靈輝，邱曉豔，王科宇（2019）以大學生這一主要城市租房群體為研究对象進行外業調查，運用 Logistic 模型對問卷調查資料進行實證分析，研究發現: 大學生租住集體建設用地建設租賃性住房意願與城市等級、城市房價、家庭收入、房屋戶型、房屋距工作地距離這五項因素呈負相關，與是否為試點城市、城鎮化率、工作狀態這三項因素呈正相關 (劉靈輝, 邱曉豔 & 王科宇, 2019)。

陳珍珠（2020）將鄭州市住房租賃市場供給主體劃分為三大類，包括個體房東、住房租賃機構和租客，通過 Probit 模型和 Logit 模型，得出年齡、受教育程度、住房租賃平臺的完善程度、租賃備案手續的繁雜程度、租賃市場監管的嚴格程度等因素會對私人住房租賃需求者產生正向影響，而租賃仲介的參與、對租賃政策的認可程度會對其租賃意願產生負向影響 (陳珍珠, 2020)。

（3）文獻評述與總結

從上述文獻可以看出，目前已存在大量關於租房意願的研究。由於國外發展較早、擁有更為成熟的租房市場，因此無論是從深度還是廣度上，對租房意願的方法論（如 Logit、Probit 模型等）以及理論和實證研究都具有較為全面的研究成果。而國內對於租房意願的研究起步較晚，所做的研究大多是在國外學者較為成熟的研究基礎上結合國內的實際情況來進行的。不同的學者在研究變數和研究樣本的選取上有所差異，但是在方法論層面，大部分學者仍以 Logit、Logistic 和 Probit 模型為主。並且，對於租房意願的研究，國內外學者大多倾向于构造解釋性模型，极少会选择去构造与租房意願相关的預測性模型。

第三章 相關算法介紹

3.1 特徵篩選三大方法

3.1.1 過濾法 (Filter)

過濾法是常用的一類特徵篩選方法，也是操作相對簡單的方法。過濾法的核心思想是：在篩選特徵時，將特徵篩選和模型的建立完全隔離開，不考慮具體使用什麼模型，只看不同的特徵對於目標變數的影響大小。

這種方法包括：移除低方差特徵、卡方檢驗、皮爾森相關係數、費雪分數等。在本文中，由於我們採用的變數即有離散型隨機變數，又有連續型隨機變數，所以移除低方差特徵並不適用。我們在單變數特徵選擇中，選擇最常用的 Pearson 相關係數作為篩選特徵的方法。

3.1.1.1 皮爾森相關係數

皮爾森相關係數也稱皮爾森積矩相關係數，是最常用的一種相關係數。它專注於兩個變數之間的相關性的研究，即探討兩個變數之間的關係密切到了什麼程度。這裡對於關係，我們給出一個定量的指標：即兩變數的實際關係接近線性關係的程度。所以皮爾森相關係數是一種線性相關係數。

總體相關係數定義為兩個變數之間的協方差與兩者標準差乘積的比值，公式如下：

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

樣本相關係數（即樣本皮爾森相關係數）常用 r 表示，定義為兩個變數的樣本之間的協方差和兩樣本標準差乘積的比值：

$$r = \frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$$

皮爾森相關係數反映了兩個變數的線性相關性程度，r 值介於-1 到 1 之間，r 的絕對值越大說明相關性越強。

當 r>0時，表明兩個變數正相關，即一個變數值越大則另一個變數值也會越大；當 r<0 時，表明兩個變數負相關，即一個變數值越大則另一個變數值反而會越小；當 r=0 時，表明兩個變數沒有線性關係，（但是可能存在其他方式的相關性，比如曲線方式）；當 r=1 和-1 時，意味著兩個變數有線性關係。

3.1.2 包裹法 (Wrapper)

不同於過濾式特徵篩選法，其篩選過程完全不考慮後續學習器，包裹式篩選法恰恰是利用最終學習器的性能來作為特徵篩選的準則。我們所做的嘗試包括並不局限於：遞迴特徵消除(Recursive Feature Elimination)、序列特徵選擇(Sequential Feature Selection Algorithms)、遺傳演算法(Genetic Algorithms)以及LVW (Las Vegas Wrapper)。

3.1.2.1 遞迴特徵消除（Recursive Feature Elimination）

遞迴特徵消除演算法是典型的包裹式演算法。我們通常用所有的特徵先跑一個模型，根據所選模型相關的係數（如準確率），刪除尾部 5%-10%的對於模型貢獻程度較小的弱特徵。接著，用剩下的特徵重新訓練模型，根據模型相關的係數，繼續刪除尾部 5%-10%的弱特徵。不斷重複上述過程，當模型相關的係數（如準確率）出現大幅的下降時，停止刪除所選特徵。

因此，遞迴特徵消除的缺點是它的穩定性完全取決於模型的穩定性，如果模型不穩定，那麼遞迴特徵消除的效果就會不穩定。

3.1.2.2 LVW（Las Vegas Wrapper）

由於 LVW 的原理在於隨機選擇特徵組合，通過分類器的表現來評判特徵組合的好壞，而我們的特徵量高達 1500 以上，LVW 的耗時相當之長。除此之外，最佳組合的特徵總數大概是多少也是未知數。LVW 採用了隨機策略，而每次特徵子集評價都需訓練學習器，計算開銷很大，因此若有執行時間限制，我們大概率是得不到解的。

3.1.2.3 遺傳演算法（Genetic Algorithms）

遺傳演算法思想起源于達爾文的進化思想，其演算法是啟發式的，故可以尋找到全域最優解，但代價是收斂速度慢，控制變數較多。基於遺傳演算法做特徵篩選是一種比較創新的想法。該演算法是以分類器的識別率作為特徵選擇的可分性判斷依據。在遺傳演算法中，對所選擇的特徵用[0,1]二進位串來初始化，由於二進位數字{0，1}是等概率出現的，所以最優特徵個數的期望是原始特徵個數的一半。要進一步減少特徵個數，則可以讓二進位數字{0，1}以不等概率出現，以 a 個特徵中選擇 b 個特徵為例，使得在 a 位元二進位串中 1 出現的概率為 b/a。其具體步驟包括：基因編碼，種群初始化，選擇操作，交叉操作，變異操作，結束條件判斷等。其中，基因操作包括選擇，交叉和突變將重複進行，以使每個樣本的 CV 分數都變得越來越高。具體流程見附錄圖 1。

3.1.3 嵌套法 (Embedding)

3.1.3.1 嵌模式

嵌入式特徵選擇法結合了過濾式和包裹式的優點，將特徵選擇嵌入到模型構建的過程中 (Zhao, Chen, Pedrycz & Wang, 2019)。在嵌入過程中，該方法會先通過一些特殊的模型擬合數據，然後根據模型內某些對於特徵評價的屬性來作為評價指標，最後再使用包裹式的特徵選擇方法來進行選擇。本文採用隨機森林和 ℓ1 正則化兩種嵌入式特徵選擇方法對特徵進行選擇。

3.1.3.2 隨機森林

隨機森林(Random Forest)是一種以決策樹為基學習器的集成學習演算法 (Breiman, 2001)。如附錄圖 2 所示，該演算法先通過自主抽樣法(Bootstrap)從訓練集中抽出 n 組樣本，通過訓練每組樣本得到對應的 n 組決策樹（本案例中決策樹為分類樹），最後由所有決策樹組成的隨機森林對測試樣本進行預測，並用票選法決定預測的結果。

為了進行有效的特徵篩選，我們可以計算出每個特徵在隨機森林中的每棵分類樹上的貢獻，然後對每個特徵在所有樹的貢獻度取平均值，最後比較特徵之間的貢獻度大小。常見的貢獻度衡量指標有三種演算法：ID3、C4.5 和 CART。

對於 ID3 演算法，假設存在一個屬性的可取值數目較多，且這個屬性對應的可取值下的樣本數量很少時，ID3 演算法下該屬性的資訊增益是過高的，即該屬性純度過高，導致這個屬性被判定為是適合劃分。因為用較多取值的屬性來進行劃分會導致決策樹的泛化能力較弱，不能夠對新樣本進行有效的預測。

CART 決策樹是使用基尼係數 (Gini Index)來選擇劃分屬性的(Breiman, 1894)。假設當前樣本集合D中第 k 類樣本所占的比例為 p_k(k = 1,2,…,|y|)，則基尼係數的定義為：

$$\text{Gini}(D) = \sum_{k=1}^{|y|} \sum_{k' \neq k} \widehat{p_k}\, \widehat{p_{k'}}$$

$$= \sum_{k=1}^{k} \widehat{p_k}(1\text{-}\widehat{p_k})$$

根據式(3.1.3.1)，Gini(D)表示的是：對資料集D進行兩次無放回抽樣，抽取的樣本中類別不一樣的概率。即用 Gini(D)表示資料集D的純度時，CART 決策樹是二叉的，即子結點個數是兩個。

而 C4.5 決策樹相對於 ID3 演算法採用增益率(Gain Ratio) 而非資訊增益(Information gain)作為屬性的劃分標準，從而避免了資訊增益準則中對可取值數目較多屬性的偏好，減少了這種偏好導致決策樹泛化能力下降的不利影響 (Wan

L, et al., 2013)。同時，C4.5 演算法是基於資訊熵(Information Entropy)對屬性的純度進行量化的，採用與式(3.1.3.1)相同的符號，資訊熵可表示為：

$$\text{Ent}(D) = -\sum_{k=1}^{|y|} \widehat{p_k} \log_2 \widehat{p_k}$$

對比式(3.1.3.1)和(3.1.3.2)，資訊熵作為純度的標準時，相對於基尼係數可以產出多個子結點個數。本案例中所需分類的標籤為租房意願的強烈程度（高、中、低），因此 C4.5 演算法相對於 CART 演算法更合適；同時 C4.5 避免了 ID3 演算法中對可取值數目較多屬性的偏好，因此在本案例中我們選擇 C4.5 演算法作為隨機森林中特徵的貢獻度衡量標準。

採用與式(2.3.3.1)相同的符號，假設離散屬性 θ 有 N 個可能的取值$\theta^+ = \{\theta^1, \theta^2, \theta^3, \theta^4, \ldots, \theta^i, \ldots, \theta^N\}$，若使用 θ 來對樣本集$D$進行劃分，則會產生 N 個分支結點，其中第 n 個分支結點包含了D中所有在屬性 θ 上取值為θ^n的樣本，記為D^n。結合(2.3.3.2)中D^n的資訊熵，考慮到不同分支結點所包含樣本數量的不同，我們對每一個分支結點賦予對應的權重$|D^n|/|D|$，從而可以計算出屬性 θ 對樣本集 D 進行劃分所得到的資訊增益(Information Gain)：

$$\text{Gain}(D,\theta) = \text{Ent}(D) - \sum_{n=1}^{N} \frac{|D^n|}{|D|} \text{Ent}(D^n)$$

為了避免 Gain(D,θ)對可取值數目較多屬性的偏好，我們引入屬性 θ 的固有值(Intrinsic Value) (Breiman, 1984)，即每個分支結點的資訊熵 Ent($|D^n|/|D|$)：

$$\text{IV}(\theta) = -\sum_{n=1}^{N} \frac{|D^n|}{|D|} \log_2 \frac{|D^n|}{|D|}$$

由式(2.3.3.4)可得，當屬性 θ 的可能取值越大，則 IV(θ)會越大，即可以通過 IV(θ)對資訊增益進行限制，從而產生新的屬性劃分指標增益率(Gain Ratio)：

$$\text{Gain Ratio}(D,\theta) = \frac{\text{Gain}(D,\theta)}{\text{IV}(\theta)}$$

$$\theta^* = \arg\max_{\theta \in \theta^+} \text{Gain Ratio}(D, \theta)$$

並且由式(3.1.3.5)的增益率可得，C4.5 演算法下決策樹的最優劃分屬性為θ^*，即使得劃分後增益率最大的屬性。

此時我們即可用增益率對每個特徵在隨機森林下的貢獻度進行衡量。我們將特徵貢獻度評分(Features Contribution Measures)用FCM來表示，計算特徵θ^i的貢獻度評分FCM_i，即第 i 個特徵在隨機森林所有決策樹中結點分裂的增益率的平均改變量。則θ^i在決策樹 j 的重要性，即決策樹 j 分枝後的增益率變化量為：

$$FCM_{ij} = \Delta\text{Gain Ratio}(D^j,\theta^i)$$

假設隨機森林中存在 m 棵樹，則θ^i的貢獻度評分FCM_i可表示為：

$$FCM_i = \frac{1}{m}\sum_{j=1}^{m} FCM_{ij}$$

通過式(3.1.3.8)的貢獻度評分FCM_i對本案例中所有特徵進行評分。

3.1.3.3 $\ell 1$ 正則化

正則化是機器學習中控制模型複雜度的方法，通過約束參數的範數來減小模型在訓練集上的過擬合現象 (Quinlan, 1993)。正則化將訓練集中每個特徵權重所對應的範數加入到模型的損失函數作為懲罰項，通過最小化損失函數，從而減少了每個特徵所對應係數的權重，降低了模型的複雜度。

訓練集中特徵權重的範數主要分為 ℓ1 範數和 ℓ2 範數，假設離散屬性 α 有 M 個可能的取值$A = \{\alpha^1, \alpha^2, \alpha^3, \alpha^4, \ldots, \alpha^j, \ldots, \alpha^M\}$，則 α 的 ℓ1 範數和 ℓ2 範數可表示如下：

$\ell 1$ 範數： $\|\alpha\|_1 = \sum_{m=1}^{M}|\alpha|$

$\ell 2$ 範數： $\|\alpha\|_2 = \sqrt{\sum_{m=1}^{M}\alpha^2}$

假設範數對應的損失函數為$\mathcal{L}(\alpha)$，正則化係數為λ，則 ℓ1、ℓ2 正則化的目標損失函數分別為：

$\ell 1$ 正則化損失函數： $\mathcal{L}(\alpha)^{(\ell 1)} = \|\alpha\|_1$

$\ell 2$ 正則化損失函數： $\mathcal{L}(\alpha)^{(\ell 2)} = \frac{1}{}\|\alpha\|_2^{\ 2}$

通過最小化損失函數，可以求出 ℓ1、ℓ2 正則化中的特徵權重α。本文通過梯度下降法(Gradient Descent)對損失函數求最小值。

分別對$\mathcal{L}(\alpha)^{(\ell 1)}$和$\mathcal{L}(\alpha)^{(\ell 2)}$求導：

$\frac{d\mathcal{L}(\alpha)^{(\ell 1)}}{d\alpha} = sign(\alpha)$, 其中$sign(\alpha) = (\frac{\alpha_1}{|\alpha_1|}, \frac{\alpha_2}{|\alpha_2|}, \frac{\alpha_3}{|\alpha_3|}, \ldots, \frac{\alpha_1}{|\alpha_1|})$

$\frac{d\mathcal{L}(\alpha)^{(\ell 2)}}{d\alpha} = \alpha$

根據式(3.1.3.13)和(3.1.3.14)可做出 ℓ1 和 ℓ2 正則化損失函數及其導數圖像如附錄圖 3-6 所示。其中圖 3 和圖 4 為 ℓ1 正則化損失函數及其導數圖像，圖 5 和圖 6 為 ℓ2 正則化損失函數及其導數圖像。

對於 ℓ1 正則化的損失函數，設學習速率為η，則根據梯度下降法可得：

$$\alpha_{new}{}^{(\ell 1)} = \alpha_{old}{}^{(\ell 1)} - \eta \cdot \frac{d\mathcal{L}(\alpha)^{(\ell 1)}}{d\alpha}$$

令α>0，根據式(3.1.3.15)，則$\frac{d\mathcal{L}(\alpha)^{(\ell 1)}}{d\alpha}$ = 1。因此，由式(3.1.3.15)可得：

$$\alpha_{new}{}^{(\ell 1)} = \alpha_{old}{}^{(\ell 1)} - \eta \cdot 1$$

當學習速率η不會因過大而無法收斂時，存在某些特徵的權重α會隨著 ℓ1 正則化損失函數最小化，而使得α梯度下降到 0，如附錄圖 7 所示。

當α<0 時，同理可得 ℓ1 正則化可以使訓練集中某些特徵的權重α變為 0，從而實現篩選變數、減少模型複雜度的效果。

對 ℓ2 正則化的損失函數進行最小化，採用與式(3.1.3.15)相同的符號，根據梯度下降法可得：

$$\alpha_{new}{}^{(\ell 2)} = \alpha_{old}{}^{(\ell 2)} - \eta \cdot \frac{d\mathcal{L}(\alpha)^{(\ell 2)}}{d\alpha}$$

把式(3.1.3.14)代入式(3.1.3.17)可得：

$$\alpha_{new}{}^{(\ell 2)} = \alpha_{old}{}^{(\ell 2)} - \eta \cdot \alpha_{old}{}^{(\ell 2)} = (1 - \eta) \cdot \alpha_{old}{}^{(\ell 2)}$$

當 0< η<1 時，由式(2.3.3.18)可知，在 ℓ2 正則化損失函數梯度下降更新權重的過程中，所有特徵的權重都無法變為 0。而當η>1 時，模型存在因為學習速率過大而無法收斂的風險。因此，ℓ2 正則化通常可以保留所有的特徵變數，減少特徵變數所對應的權重，從而實現減少過擬合的效果。

由於本案例是分類問題，且在特徵構造部分曾構建了大量的稀疏資料，使用 ℓ2 正則化無法起到減少篩選變數或減少模型複雜度的作用，因此我們選擇 ℓ1 正則化作為篩選變數的方法。

對於 ℓ1 正則化，假設訓練集的樣本數量為 N，損失函數為$\mathcal{L}(x, \alpha)$，待學習的模型為$f_\alpha(x)$，真實值為y，正則化係數為λ，則 ℓ1 正則化的目標損失函數為：

$$\mathcal{L}(x, \alpha)^{(\ell 1)} = \frac{1}{N}\sum_{i=1}^{N}(f_\alpha(x^{(i)}) - y^{(i)})^2 + \lambda\|\alpha\|_1$$

使用梯度下降法對 ℓ1 正則化的損失函數進行最小化：

$$\alpha_0 = \alpha_0 - \eta \cdot \frac{\partial \mathcal{L}(x,\alpha)^{(\ell 1)}}{\partial \alpha_0}$$

$$= \alpha_0 - \eta \cdot \frac{1}{N}\sum_{i=1}^{N}(f_\alpha(x^{(i)}) - y^{(i)})\, x_0{}^{(i)}$$

$$\alpha_j = \alpha_j - \eta \cdot \frac{1}{N}\sum_{i=1}^{N}(f_\alpha(x^{(i)}) - y^{(i)})\, x_j{}^{(i)} - \underset{(j=1,2,3,\ldots,n)}{\frac{\lambda}{N}\alpha_j}$$

$$= \alpha_j(1 - \eta\frac{\lambda}{N}) - \eta \cdot \frac{1}{N}\sum_{i=1}^{N}(f_\alpha(x^{(i)}) - y^{(i)})\, x_j{}^{(i)}$$

綜合式(3.1.3.20)和式(3.1.3.21)，訓練集中特徵的權重α可表示為：

$$\alpha = \begin{cases} \alpha_0 - \eta \cdot \frac{1}{N}\sum_{i=1}^{N}\left(f_\alpha\left(x^{(i)}\right) - y^{(i)}\right) {x_0}^{(i)}, \alpha = \alpha_0 \\ \alpha_j\left(1 - \eta\frac{\lambda}{N}\right) - \eta \cdot \frac{1}{N}\sum_{i=1}^{N}\left(f_\alpha\left(x^{(i)}\right) - y^{(i)}\right) {x_j}^{(i)}, \alpha = \underset{(j=1,2,3,\ldots,n)}{\alpha_j} \end{cases}$$

根據式(3.1.3.22)，通過調節學習速率η，即可得到每個特徵的權重α以及經過篩選後留下的特徵變數。

3.2 欠採樣過採樣

3.2.1 採樣法原理

本案例中，我們所需預測的因變數“興趣程度”呈現出類別不平衡的現象。如圖 8 所示，在某單一房源的租房興趣意願方面，興趣程度的高、中、低比例大約為 7:2:1，即不同興趣程度的訓練樣本數目差別很大。

因變數的類別不平衡容易導致分類器的分類結果偏向樣本數目較多的類（即興趣程度低的一類），從而影響分類效果。因此，我們需要通過欠採樣和過採樣技術使因變數的類別達到數量平衡的狀態。

3.2.1.1 欠採樣

欠採樣(Undersampling)是一種緩解類不平衡的方法，其通過拋棄樣本的方式來實現，即對訓練集內樣本數量較多的類別進行少採樣(Koziarski, 2020)。在欠採樣技術中，我們一般是在保留少數類樣本不變的基礎上，預設出多數類與少數類樣本的目標數量比例，通過無放回地隨機抽樣來剔除多數類樣本。在本案例中，我們通過欠採樣技術將興趣程度的高、中、低比例設為 1:1:1。

3.2.1.2 過採樣

過採樣(Oversampling)和欠採樣一樣，都是緩解類不平衡的方法。與欠採樣不同，過採樣對訓練集中樣本數量較少的類別進行資料的生成，從而達到預設的多數類與少數類樣本的目標數量比例(Koziarski, 2020)。常見的過採樣技術有 5 種：樸素隨機過採樣演算法、SMOTE 演算法、ADASYN 演算法、SMOTEENN 演算法和 SMOTETOMEK 演算法。本文將詳細討論這 5 種演算法，並通過模型預測結果對這 5 種演算法進行評價。

3.2.2 欠採樣過採樣相關演算法

3.2.2.1 樸素隨機過採樣演算法

針對不平衡資料, 樸素隨機過採樣演算法(Random Over Sampling)是從少數類的樣本中進行隨機採樣來增加新的樣本(Chawla, Hall, Bowyer & Kegelmeyer, 2002)。

但是，樸素隨機過採樣把少數類樣本複製多份後，複製的樣本會在高維空間中反復出現，訓練出來的模型會存在一定的過擬合。為了解決這一問題，我們在每次生成新資料點時加入了輕微的隨機擾動以防過擬合。

3.2.2.2 SMOTE 演算法

SMOTE 演算法(Synthetic Minority Oversampling Technique)是對少數類樣本進行分析並根據少數類樣本人工合成新樣本添加到資料集中的一種過採樣演算法(Chawla, Hall, Bowyer & Kegelmeyer, 2002)。該演算法結合了 KNN 演算法對少數類樣本進行了類比，詳細流程如下：

1.假設少數類樣本$X_S = \{x_{s1}, x_{s2}, x_{s3}, x_{s4}, \ldots, x_{si}, \ldots, x_{sn}\}$，對於少數類中每一個樣本$x_{si}$，計算該樣本與少數類中其他樣本的歐式距離，得到最近的k個近鄰（即對少數類點進行 KNN 演算法）。

2.根據樣本不平衡比例設置一個採樣比例以確定採樣倍率（本文將採樣比例設置為 1:1:1），對於每一個少數類樣本x_{si}，從其k近鄰中隨機選擇若干個樣本。

3.假設選擇的近鄰為x_{si}'，對於每一個隨機選出的近鄰x_{si}'，分別與原樣本按照如下的公式構建新的樣本：

$$x_{new} = x_{si} + rand(0,1) \cdot (x_{si}' - x_{si})$$

雖然 SMOTE 演算法從一定程度上通過 KNN 演算法減少了樸素隨機過採樣演算法的過擬合問題，但是由於該演算法對每個少數類樣本都生成新樣本，因此容易發生生成樣本重疊(Overlapping)的問題，增加了類之間重疊的可能性；另外，該演算法可能會生成一些沒有提供有益資訊的樣本。

3.2.2.3 ADASYN 演算法

ADASYN 演算法名為自我調整合成抽樣 (Adaptive Synthetic Sampling)，其最大的特點是採用某種機制自動決定每個少數類樣本需要產生多少合成樣本，而不是像 SMOTE 那樣對每個少數類樣本合成同數量的樣本，從而避免了不必要的類別的重疊性(Chawla, Hall, Bowyer & Kegelmeyer, 2002)。

少數類樣本的表示採用與式(3.2.2.1)相同的符號，假設多數類樣本$X_L = \{x_{l1}, x_{l2}, x_{l3}, x_{l4}, \ldots, x_{li}, \ldots, x_{ln}\}$，樣本類別的不平衡度為$\delta$，預設的採樣比例為$\delta_0$，需要合成的樣本數量為$\Delta X$，先對樣本類別的不平衡度進行計算：

$$\underset{\delta\in(0,1]}{\delta} = X_S/X_L$$

當$\delta<\delta_0$時，對需要合成的樣本總量進行計算：

$$\Delta X = (X_L\text{-}X_S)\beta$$

其中，式(3.2.2.3)的$\beta \in [0,1]$，用來調節所需合成的樣本數量以達到預設的目標採樣比例δ_0。

依據所設定的ΔX，同 SMOTE 演算法流程的第一步，對每個屬於少數類的樣本x_{si}用歐式距離計算k_i個近鄰。假設k_i'為k_i個鄰居中屬於多數類的樣本數目，則k個近鄰中多數類樣本對總體樣本的比例r_i為：

$$r_i = k_i'/k_i$$

對 KNN 演算法中計算出的每個樣本k近鄰中多數類樣本對總體樣本的比例r_i進行標準化，以便標準化後的該比例後續作為每一個樣本的權重，針對個體所處多數類樣本數量不同的環境合成不同數量的樣本。

$$\widehat{r_i} = r_i/\sum_{i=1}^{n} r_i$$

式(3.2.2.5)標準化後得到的$\widehat{r_i}$滿足分佈$\sum_{i=1}\widehat{r_i} = 1$。

通過$\widehat{r_i}$對每一個少數類的樣本x_{si}所需合成樣本的數量進行計算：

$$\Delta x_i = \widehat{r_i}\Delta X$$

在每個待合成的少數類樣本周圍k近鄰中選擇1個少數類樣本x_{ki}，根據下列等式進行合成樣本x_{si}''：

$$x_{si}'' = x_{si} + (x_{ki} - x_{si}) \cdot \tau$$

式(2.1.2.7)中τ為隨機擾動因數，$\tau \in [0,1]$。重複式(3.2.2.4) 到 (3.2.2.7)，合成樣本直至$\delta=\delta_0$或滿足ΔX為止。

ADASYN 演算法針對不同少數類樣本合成不同數量的樣本，解決了 SMOTE 演算法中可能存在的類別重疊性，但是該演算法易受離群點的影響，如果一個少數類樣本的k近鄰都是多數類樣本，則其權重會變得相當大，進而會在其周圍生成較多的樣本。

3.2.2.4 SMOTEENN 演算法

SMOTEENN 演算法是對於 SMOTE 演算法的改良，在 SMOTE 演算法的基礎上，KNN 演算法作為監督機制對 SMOTE 演算法的合成結果進行篩選。對於合成的少數類樣本，如果其k近鄰中有超過一半都不屬於多數類，則這個樣本會因為類別不符而被剔除。相對於 SMOTE 演算法，該衍生演算法利用 KNN 演算法作為監督機制糾正了合成資料中的分類錯誤。

3.2.2.5 SMOTETOMEK 演算法

SMOTETOMEK 演算法結合了 SMOTE 演算法和 Tomek link 數對。假設少數類樣本中的樣本點x_{si}和x_{sj}屬於不同的類，如果x_{si}的k最近鄰（即$k = 1$時）有且只有x_{sj}，且x_{sj}的k最近鄰也有且只有x_{si}，那麼 (x_{si}, x_{sj})被稱為一組 Tomek link 數對。與 SMOTEENN 演算法類似，Tomek link 數對作為監督機制對 SMOTE 演算法的合成結果進行篩選。對於合成的少數類樣本，如果存在兩個樣本點為 Tomek link 對，則其中某個樣本可視為雜訊(偏離正常分佈太多)而被剔除。

3.3 Boosting 及相關算法

3.3.1 集成學習

近年來，集成學習（Ensemble Learning）在 Kaggle 等各大資料競賽平臺中屢屢得到不錯的成績，漸漸引起大家的關注。它的中心思想主要關注於弱學習器的集成，俗稱“三個臭皮匠，頂一個諸葛亮”。若每個學習器的種類一樣，則可稱為“基學習器”（Base Learner）；若不一樣，則稱為“組件學習器”（Component Learner）。學習器需要滿足以下兩個條件才能最終得到一個強學習器：第一，學習器之間要有差異；第二，每個學習器的性能不可太差。當我們對相對較弱的學習器給予更多的關注時，強學習器的整體性能將會不斷提升。

總體而言，集成學習分為兩大類：一種是以 Boosting 為代表的，基學習器相互依賴的，串列生成的方法，另一種則是以 Bagging 與隨機森林為代表的，基學習器不存在相互依賴關係的，並行生成的方法。Bagging 採用均勻取樣，而 Boosting 根據錯誤率來取樣，因此 Boosting 的分類精度要優於 Bagging。本文所用的主要模型全部都基於 Boosting 演算法， 涉及模型包括 AdaBoost, GBDT, 與 XGBoost。以下將對每一個演算法進行推算介紹。

3.3.2 Boosting 算法

Boosting 族可通過給予弱學習器不同的權重組合成強學習器，其演算法原理為：先從初始訓練集當中訓練出一個基學習器，並根據其效能對訓練樣本分佈進行調整，即對前基學習器判斷錯誤的樣本增加關注度，再利用調整後的樣本來訓練下一個基學習器；如此往復，直至學習器數量達到預先設定值 T，將 T 個基學習器進行加權融合得到強學習器。從偏差-方差 (Bias-Variance) 的角度來看，相比於 Bagging, Boosting 的目的更加關注於降低偏差，因此 Boosting 能基於泛化性較弱的學習器來構造出很強的學習器。

3.3.3 AdaBoost 三分類問題

AdaBoost（Adaptive Boosting）演算法是 Boosting 族演算法中最著名的代表，由 Freund and Schapire 於 1997 年提出。AdaBoost 最常見也是最易於理解的推導方式基於"加性模型" (additive model), 即通過基學習器的線性組合來最小化指數損失函數 (exponential loss function) 。AdaBoost 的基本演算法頻繁地應用於二分類場景，然而我們需要處理的是三分類問題。因此我們將引入 SAMME 演算法 (Hastie, Rosset, Zhu & Zou, 2009)，接下來我們會通過對比，簡短地介紹一下二者的區別：AdaBoosting 主要關注於兩個問題：第一，如何更新每個模型的權重；第二，如何更新樣本分佈，即給予判斷錯誤的樣本更多的關注。在計算 α(模型權重)時，相較於二分類，三分類問題將會引入$\log(K-1)$，具體原因我們將根據 SAMME 演算法 (Stagewise Additive Model using a Multi-class Exponential loss function) 展開介紹：

1. 標籤和輸出的向量化

SAMME 演算法是 AdaBoost 的推廣，其中X_i是樣本的特徵向量，y_i是樣本的標籤向量，$y_i = (y_{i1}, y_{i2}, y_{i3}, \dots, y_{iK})$，其分量定義如下：

$$y_{ik} = \begin{cases} 1, & \text{如果}\mathrm{x_i}\text{属于第 k 类} \\ -\dfrac{1}{K-1}, & \text{如果}\mathrm{x_i}\text{不属于第 k 类} \end{cases}$$

例如，假設訓練集第一個樣本$(\mathrm{x_I}, y_i)$屬於第一類，那麼$y_1 = \left(1, -\frac{1}{K-1}, -\frac{1}{K-1}, \dots, -\frac{1}{K-1}\right)$，在這個演算法中分類器依然採用加法模型，我們將訓練好的分類器記為$f_m(x)$, 那麼它可以被寫作以下形式：

$$f_m(x) = \sum_{i}^{m} \beta_i \, \boldsymbol{g_i}(\boldsymbol{x})$$

其中， $g_i(x)$ 是基分類器(弱分類器)，其輸出也是向量，為了保證求解得到的 $f_m(x)$唯一，我們可以給它加上一個對稱的約束條件，限制$f_m(x)$各分量之和為 0，即 $f_{m1}(x) + f_{m2}(x) + \cdots + f_{mK}(x) = 0$ 。由於 $f_m(x)$ 可表示成 m 個基分類器之和，且 $f_m(x)$ 的輸出是向量，故 $g_i(x)$ 的輸出也應為向量，並也滿足對稱的約束條件，即 $g_{m1}(x) + g_{m2}(x) + \cdots + g_{mK}(x) = 0$ 。 最終我們得到的 $\boldsymbol{y}$ 是包含 K 個 K 維向量的集合:

$$\boldsymbol{y} = \left\{ \begin{matrix} \left(1, -\frac{1}{K-1}, \ldots, -\frac{1}{K-1}\right)^{\boldsymbol{T}}, \\ \left(-\frac{1}{K-1}, 1, \ldots, -\frac{1}{K-1}\right)^{\boldsymbol{T}}, \\ \vdots \\ \left(-\frac{1}{K-1}, \ldots, -\frac{1}{K-1}, 1\right)^{T} \end{matrix} \right\}.$$

2. 模型求解

該演算法的損失函數沿用了傳統 AdaBoost 的指數損失函數，標量需要轉化成向量，得到推廣到多類情況的指數損失函數如下：

$$L\big(y, f(x)\big) = \exp\left(-\frac{1}{K}(y1f1 + \cdots + yKfK)\right) = \exp\left(-\frac{1}{K}\, y^T f\right)$$

為了後面計算方便，作者 Ji Zhu 等人引入了一個常數$\frac{1}{K}$ 。和 Adaboost 演算法的推導一樣，我們只關注當前輪的訓練，因此 $f_m(x) = f_{m-1}(x) + \beta_m \, \boldsymbol{g_m}(x)$，上一輪訓練的結果 $f_{m-1}(x)$ 被吸收進樣本權重項：

$$\begin{aligned} L\big(y, f(x)\big) &= \exp\left(-\frac{1}{K}(y1f1 + \cdots + yKfK)\right) \\ &= \sum_i exp\left(-\frac{1}{K}\, y_i * f_m(x_i)\right) \\ &= \sum_i exp\left(-\frac{1}{K}\, y_i * \left(f_{m-1}(x_i) + \beta_m \, \boldsymbol{g_m}(x_i)\right)\right) \\ &= \sum_i w_i \, exp\left(-\frac{\beta_m}{K}\, y_i * \boldsymbol{g_m}(x_i)\right) \end{aligned}$$

經過簡單的運算，預測正確與預測錯誤的兩種情況下的內積結果為：

$$\begin{cases} \text{預測正確：} & y_i * \boldsymbol{g_m}(x_i) = 1 + \left(-\frac{1}{K-1}\right)^2 (K-1) = \frac{K}{K-1} \\ \text{預測錯誤：} & y_i * \boldsymbol{g_m}(x_i) = -\frac{2}{K-1} + \left(-\frac{1}{K-1}\right)^2 (K-2) = \frac{K}{(K-1)^2} \end{cases}$$

將以上結果帶入損失函數運算式可得：

$$\boldsymbol{L(y, f(x))} = exp\left(-\frac{\beta_m}{K-1}\right)\sum_i w_i + \left(exp\left(-\frac{\beta_m}{(K-1)^2}\right) - exp\left(-\frac{\beta_m}{K-1}\right)\right)\sum_i w_i I(\boldsymbol{g_m} \neq y_i)$$

因此我們令損失函數對β_m的導數等於 0 可得：

$$\frac{\partial L}{\partial \beta_m} = \left(-\frac{1}{K-1}\right) exp\left(-\frac{\beta_m}{K-1}\right)\sum_i w_i + \left(\frac{1}{(K-1)^2}\ exp\left(\frac{\beta_m}{(K-1)^2}\right) + \frac{1}{K-1}\ exp\left(-\frac{\beta_m}{K-1}\right)\right)\sum_i w_i I(\boldsymbol{f_m} \neq y_i) = 0$$

即，

$$\left(\frac{1}{K-1}\right) exp\left(-\frac{\beta_m}{K-1}\right)\sum_i w_i = \left(\frac{1}{(K-1)^2}\ exp\left(\frac{\beta_m}{(K-1)^2}\right) + \frac{1}{K-1}\ exp\left(-\frac{\beta_m}{K-1}\right)\right)\sum_i w_i I(\boldsymbol{f_m} \neq y_i)$$

$$exp\left(-\frac{\beta_m}{K-1}\right) = \left(\frac{1}{K-1}\ exp\left(\frac{\beta_m}{(K-1)^2}\right) + exp\left(-\frac{\beta_m}{K-1}\right)\right)\boldsymbol{r_{error}}$$

$$\frac{1-\boldsymbol{r_{error}}}{\boldsymbol{r_{error}}} = \frac{1}{K-1}\ exp\left(\frac{K\beta_m}{(K-1)^2}\right)$$

$$\beta_m = \frac{(K-1)^2}{K}\left(\log\left(\frac{1-\boldsymbol{r_{error}}}{\boldsymbol{r_{error}}}\right) + \log(K-1)\right)$$

其中，$\boldsymbol{r_{error}} = \frac{\sum_i w_i I(\boldsymbol{f_m} \neq y_i)}{\sum_i w_i}$，代表模型的錯誤率；接下來便得到了 SAMME 演算法。

3. SAMME 演算法

1) 初始化觀測值的權重 $w_i = \frac{1}{n}, i = 1,2,\ldots,n.$

2) 對於分類器 $\boldsymbol{g_m}(x_i), m = 1,2,\ldots,M.$；在權重$w_i$下訓練分類器 $\boldsymbol{g_m}(x_i)$

(a) 使用權重w_i將分類器 $\boldsymbol{g_m}(x_i)$擬合到訓練數據。

(b) 計算分類器錯誤率:

$$\boldsymbol{r_{error}} = \frac{\sum_i w_i\, I(\,\boldsymbol{f_m} \neq\, y_i)}{\sum_i w_i}$$

(c) 計算模型權重:

$$\alpha_m = \log(\frac{1 - \boldsymbol{r_{error}}}{\boldsymbol{r_{error}}}) + \log(K-1)$$

(d) 更新樣本權重w_i ：

$$w_i = w_i \exp\big(\alpha_m I(\,\boldsymbol{g_m} \neq\, y_i)\big)$$

$$i = 1,2,\ldots,n.$$

(e) 樣本權重w_i歸一化

3) 轉到步驟 2，直到錯誤率小於閾值

4) 輸出：

$$F(x) = \arg\max_k \sum_{m=1}^{M} \alpha_m *\; \boldsymbol{I}(\,\boldsymbol{g_m}(x_i) = k\,)$$

3.3.4 GBDT 算法

GBDT 算法 (Gradient Boosting Decision Tree) 是通过采用加法模型（即基函数的线性组合），以及不断减小训练过程产生的残差来达到将数据分类或者回归的一种算法(Weng & Xiang, 2020)。GBDT 算法由 Gradient Boosting Modeling (GBM) 和决策树(Decision Tree)组成。对于 GBM，依据梯度下降法的不同，GBM 存在三种类型：BGBM (Batch Gradient Boosting Modeling)，SGBM (Stochastic Gradient Boosting Modeling)和 M-BGBM (Mini-Batch Boosting Modeling)。而 GBDT 算法中的决策树或弱分类器使用的是 CART 树。

对于 BGBM 算法，该算法是基于批量梯度下降法(BGD)，该算法每一次的参数更新都用到了所有的训练数据。假设存在弱分类器$g_i(x)$，真实值为y_i，学习速率为η_i，其中$i \in (1, n)$；训练集的自变量为x_j，其中$j \in (1, m)$；训练集的样本数量为 N，损失函数为$\mathcal{L}(x)$，生成的新的强分类器为$G(x)$，则 BGBM 的目标损失函数为：

$$\mathcal{L}_i(x) = \frac{1}{N}\sum_{j=1}^{N}(\,g_i(x_j) - y_{ij})^2$$

采用 BGD 法(Batch Gradient Descent)对式(3.3.4.1)中的损失函数进行最小化：

$$G(x)_1 \;=\; g_1(x) \text{ - } \eta_1 \cdot \frac{\partial \mathcal{L}_1(x)}{\partial g_1(x)}$$

$$G(x)_i = G(x)_{i-1} - \eta_i \cdot \frac{\partial \mathcal{L}_i(x)}{\partial g_i(x)}$$

综合(3.3.4.2)和(3.3.4.3)，可求出 BGBM 生成的新的强分类器：

$$G(x) = -\sum_{i=1}^{n} \eta_i \cdot \frac{\partial \mathcal{L}_i(x)}{\partial g_i(x)} + g_i(x)$$

在式(3.3.4.4)中，对于 BGBM 生成的新的强分类器$G(x)$，$\frac{\partial \mathcal{L}(x)}{}$为组成$G(x)$的组合弱分类器，$\eta_i$为对应弱分类器的权重。

SGBM 算法不同于 BGBM 算法每次需要用到训练集中所有的样本，SGBM 算法基于随机梯度下降法(SGD)，每次迭代更新只用到一个样本。因此，SGBM 算法在式(3.3.4.1)的基础上改变了损失函数，即：

$$\mathcal{L}_i(x)' = (g_i(x_j) - y_{ij})^2$$

则 SGBM 算法生成的强分类器：

$$G(x)' = -\sum_{i=1}^{n} \eta_i \cdot \frac{\partial \mathcal{L}_i(x)'}{\partial g_i(x)} + g_i(x)$$

对于 M-BGBM 算法，该算法综合了 BGBM 和 SGBM 算法，每次迭代只使用部分样本。假设训练集分成K 组，每组数据量$N_j = \{N_1, N_2, N_3, \ldots, N_K\}$。则 M-BGBM 算法的损失函数：

$$\mathcal{L}_i(x)'' = \frac{1}{N_K} \sum_{j=1}^{N_K} (g_i(x_j) - y_{ij})^2$$

则 M-BGBM 算法生成的强分类器：

$$G(x)'' = -\sum_{i=1}^{n} \eta_i \cdot \frac{\partial \mathcal{L}_i(x)''}{\partial g_i(x)} + g_i(x)$$

3.3.5 XGBoost 算法

XGBoost (Extreme Gradient Boosting)是“極端梯度提升” 的簡稱，是一種複合樹模型，主要用於解決監督性學習問題，在分類、回歸及排序任務上都有很好的表現。與其他梯度提升演算法不同的是，該演算法可以實現平行計算、近似建樹、對稀疏資料的有效處理以及記憶體使用優化等功能。因此，在同等情況下，XGBoost 的計算速度將比同類模型快十倍以上。

1. 演算法優勢

XGBoost 的重要創新點在於，該模型通過三種方式解決了正則化問題：第一個方法為在損失函數中增加對於模型複雜度的懲罰項；第二個方法為設置步長，

以此降低每個模型的貢獻；第三個方法則與隨機森林相似，在進行行採樣的同時，也可以實現列採樣。

2. 目標函數

XGBoost 在目標函數上做出了重要創新，加入對於模型複雜度的懲罰項以實現正則化。如下式所示，在第 t 次反覆運算中，對於第 i 個樣本，我們需要在第 t-1 次反覆運算的基礎上加入$f_t(x_i)$，並使函數$L^{(t)}$達到最小值。

$$L^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y_i}^{(t-1)} + f_t(x_i)\right) + \Omega(f_t)$$

$$where\ \Omega(f) = \gamma T + \frac{1}{2}\lambda\|w\|^2$$

$L^{(t)}$中，$l\left(\mathrm{y_i}, \hat{\mathrm{y}_\mathrm{i}}^{(\mathrm{t-1})} + \mathrm{f_t(x_i)}\right)$為損失函數，描述了實際值$\mathrm{y_i}$與預測值$\hat{\mathrm{y}_\mathrm{i}}$之間的差異；$\Omega(\mathrm{f_t})$為正則項，用於控制模型的複雜度，以防止過擬合，其中w表示葉子節點上的分數所組成的向量，T表示葉子節點的數量。

一般情況下，我們可以用泰勒展開式處理損失函數，從而快速優化目標函數。

$$L^{(t)} \cong \sum_{i=1}^{n}\left[l\left(y_i, \hat{y}^{(t-1)}\right) + g_i f_t(x_i) + \frac{1}{2}h_i {f_t}^2(x_i)\right] + \Omega\left(f_t\right)$$

$$where\ g_i = \partial_{\hat{y}^{(t-1)}} l\left(y_i, \hat{y}^{(t-1)}\right)\ and\ h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}^{(t-1)})$$

在刪除常數項後，第 t 次反覆運算中的目標函數則變為：

$$L^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i {f_t}^2(x_i)\right] + \Omega\left(f_t\right)$$

此時，目標函數只需要考慮g_i和h_i，這就是 XGBoost 能夠支援自訂損失函數的原因。我們能夠優化每一個損失函數，包括邏輯回歸和加權邏輯回歸，只需要把對應的g_i和h_i作為輸入傳入即可。

3. 葉子節點權重

將正則項展開後，經過泰勒展開後的函數可以進一步整理：

$$\tilde{L}^{(t)} = \sum_{i=1}^{n}\left[g_i f_t(x_i) + \frac{1}{2}h_i {f_t}^2(x_i)\right] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2] + \gamma T$$

其中，$I_j = \{i|q(x_i) = j\}$表示每個映射到第j個葉子節點對應的資料樣本。因為$L^{(t)}$中每一個w_j都是相互獨立的，所以我們可以在當前樹的結構 q(x)已知的情況下，求出目標函數最小值所對應的w_j：

$$w_j^* = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}$$

在求得w_j^*之後，我們可以將其帶入$\tilde{L}^{(t)}$中，從而得出其最優值：

$$\tilde{L}^{(t)}(\mathrm{q}) = -\frac{1}{2}\sum_{j=1}^{T}\frac{(\sum_{i\in I_j} g_i)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T$$

$\tilde{L}^{(t)}$(q)可用於對結構為 q(x)的樹模型進行打分。該公式的計算結果類似於基尼係數，但是所適用的目標方程範圍更廣。

4. 樹結構的構建

我們將採用貪心法構建樹模型，即在樹的每個層構建的過程中，選擇當前收益最優的方案來優化目標。該過程的目標函數為分割後在目標上所獲得的收益（Gain），如下：

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma$$

該目標函數由四部分組成：如果劃分，劃分後左邊節點所帶來的收益；如果劃分，劃分後右邊節點所帶的收益；如果不劃分，原節點的收益；如果劃分，劃分後正則項的收益。

由於正則項的存在，該函數的結果不可能一直高於閾值。而當劃分後葉子節點所帶來的收益小於閾值時，我們將停止劃分；同時，我們也可以將其作為優化剪枝的依據。

第四章 數據來源及數據描述性分析

4.1 資料來源

本文的資料來自於一個由 Two Sigma 和 RentHop 共同舉辦的數據大賽，該比賽鼓勵人們利用現實世界的資料解決日常問題。數據具體來自於公寓清單網站 renthop.com，這些公寓都位於紐約市。數據中的唯一目標變數 interest_level 由在網站存在期間的查詢頻次決定。

4.2 描述性分析

4.2.1 變數類型

該資料總共有 49352 條記錄，包括 14 個引數和 1 個因變數。引數中，有六個數值類型特徵，分別為 bathroom(浴室個數）、bedrooms（臥室個數）、latitude（經度）、longitude（緯度）、listing_id（列表 id）、price（價格）；還有 8 個類別型特徵，分別為 created（資料創建日期）、building_id（房屋編號）、description（描述）、display_address(展示位址）、feature（特徵）、manager_id(經理編號）、photos（照片）、street_address（街道位址）為類別型特徵；而目標變數 interest_level(興趣程度）是類別型特徵，取值空間為 low, medium, high。

4.2.2 視覺化分析

我們首先觀察數值型特徵 price（價格）。在去除一些離群點後，可以畫出價格的密度函數（如圖 9），由圖可知，該資料集中價格呈略微右偏。

在類別型特徵(features)中，Features 變數包含了許多對房屋正面描述的關鍵字，我們做出關鍵字出現個數的柱狀圖。根據圖 10，關鍵字個數的分佈呈總體右偏，主要集中在一至六個。房屋正面資訊的多少會影響顧客的興趣程度，因此下文會對該類別型變數進行進一步處理，以便更好地訓練模型。

接著，我們對應變數-興趣程度進行視覺化。興趣程度共有三大類：低，中，高。所以之後面對的是一個三分類的預測問題。根據如下圖 11 和圖 12，興趣程度的三類數量很不均勻，興趣程度低的一類高達 69.5%，而興趣程度高的一

類僅占 7.8%。鑒於此情況，我們在建模之前會採取過採樣及欠採樣，將三類應變數的數量調至相等，以達到提高預測準確度的目的。

最後，我們將分析一下房屋的價格與租賃興趣程度的關係。從圖 13 得，當房屋租賃價格高於 8000 元時，顧客的興趣程度大多為低。相比于高價房，低價房更受大家的青睞。

第五章 數據預處理及特徵工程

5.1 數據預處理

5.1.1 缺失值處理

許多數據都存在著部分數據缺失的問題，因此我們首先檢驗該數據的完整性。我們在 python 中，調用 isnull().any(axis = 1)，發現此數據集中沒有缺失值。但是在經緯度的變量中，發現了一些顯示爲 0 的無意義的值，因此我們將這些值當成缺失值進行處理，並刪除這些無意義的值所在的行。

5.2 特徵工程

5.2.1 處理 features 變量

5.2.1.1 對 Features 變量進行疏鬆陣列處理

Features 變數包含了許多對房屋正面描述的關鍵字，是對原有特徵集很好的擴充。圖 14 是部分 features 的數據展示。

通過 for 迴圈，我們發現不同房屋關鍵字的數量大不相同，從 0 個到 39 個不等。我們認為對於房屋正面描述的關鍵字越多，顧客就會獲得更多關於該房屋的正面資訊，使用者對該房屋的興趣熱度就會提高。所以我們決定根據 features 變數建立一個疏鬆陣列，每一個關鍵字都作為疏鬆陣列中的一個變數，1 表示該條資料包含此關鍵字，0 則表示不包括。

我們首先找出 features 中所有的關鍵字，接著使用兩種方法去重，最終得到 1556 個關鍵字。最後，使用 for 迴圈得出疏鬆陣列。

5.2.1.2 处理编码类变量

本案例中存在的編碼類變數主要有 3 種：street_address，building_id 和 manager_id。以 street_address 為例，資料集一共存在 49352 個樣本，而該變數中存在多達 15358 種不同的地名。為了保證每個地名的出現次數可以作為權重充分被模型利用，我們以每種地名出現的次數作為該地名的新編碼。building_id 和 manager_id 同理，以每個地名的出現次數作為該地名的新編碼。

5.2.2 處理時間變量

通過觀察 created 列（如圖 15 所示）， 我們發現 created 是一個日期特徵，代表的是租賃資訊被創建的時間。調入 pandas 的 to_datetime， 我們將日期分割為年、月、日等不同維度。我們發現資料的創建日期集中在 2016 年的 4-6 月，為了後續更好地建模，我們僅保留了月份，並貼以數字標籤 4，5，6。結果如圖 16 所示。

5.2.3 處理地理位置變量

通過原資料，我們可以得到每條資訊的經度及緯度資訊。由於資料集中房源範圍為紐約市區，而由於紐約市五大區之間的消費水準、人中分佈等特徵並不相同，因此我們推測，在經緯度差異相同的情況下，同區資料之間的差異與跨區資料之間的差異可能會有所出入。因此，為了獲取更加詳盡的地理資訊，我們從 Google 官方網站申請 API 金鑰，並且通過 python 依據房源經緯度回溯其所屬區域，並將其更新到資料集中。結果如圖 17 所示。

第六章 基於 Boosting 算法的租房意願預測模型

6.1 特徵篩選

在對數據進行特徵工程之後，我們的特徵數量超過了 1500 個。對于這樣的高維度數據，我們需要進行特徵篩選，去除一些不相關的特徵（即噪聲），以避免因維數過高而導致一系列問題，提高對于房屋租賃意願預測的準確度。我們將分別運用過濾法、包裹法和嵌套法這三類特徵篩選方法，進行特徵的篩選工作。

6.1.1 過濾法 (Filter)

過濾法將特徵篩選與模型建立完全隔離，所以此特徵選擇的過程與學習器無關。在過濾法中，我們選擇皮爾森相關係數法對特徵進行過濾。通過尋找每一個自變量與應變量之間的相關係數 r，確定每個自變量和應變量之間綫性關係的强弱。r 的絕對值越大，相關性越强。當|r|大于等于 0.8 時，則認爲兩個變量之間高度相關，而當|r|小于 0.3 時，認爲變量之間關係極弱，甚至幾乎不相關。通過分析，我們發現在現有的 1500 多個變量中，所有資料與應變量的相關係數絕對值均低於 0.3，并且僅有六個数值大於 0.1，詳細結果如圖 18 所示。

這表明在此數據中，所有的自變量與應變量之間的綫性關係並不明顯。而皮爾森相關係數法就是通過自變量與應變量之間的綫性關係來篩選變量，所以該方法在此數據中并不適用。

6.1.2 包裹法 (Wrapper)

和過濾法不同，包裹法利用特定的學習器選擇最優的特徵子集。這裏我們選擇包裹法中最爲典型的遞回特徵消除法。由於本文要處理的是多分類問題，因此在模型上我們選擇了决策樹分類器(Decision Tree Classifier)，並選用基尼係數作為目標函數，最終通過遞回特徵消除模型篩選出了 50 個變量。

關於 LVW 與遺傳算法，我們意識到 LVW 的枚舉方式對於我們變量組合的選擇的計算成本之大是難以想象的。初次以外，相比於 RFE，LVW 的篩選並沒有條件去限制，同時，變量的個數也難以確定，因此我們在初步嘗試后放棄 LVW 而轉想遺傳算法。根據我們的遺傳算法函數的參數調整，我們分別做過兩次變量篩選，一次輸出 14 個變量，另一次為 55，在嘗試當中，我們發現大部分輸出的

變量來自稀疏矩陣，我們猜測是稀疏矩陣在一定程度上影響了遺傳算法在搜尋全局最優解的過程。

綜上，我們在 Wrapper 思想下對篩選變量的分別進行了 RFE、LVW 與遺傳算法的嘗試，然而表現不盡理想。通過文獻回顧，我們意識到梯度下降在解決此類問題表現的更爲高效，而啓發式算法在搜尋變量組合的過程中消耗了過量的時間。筆者能力受限，故轉向嵌套法繼續嘗試變量篩選。

6.1.3 嵌套法 (Embedding)

嵌套法結合了過濾法和包裹法的優點，將特徵篩選過程嵌入到模型中，先對每個特徵或屬性的重要性進行評價，後依據模型的性能決定特徵的篩選和保留數量。對於嵌套法中的隨機森林演算法依據 Boosting 模型的最終性能，我們選用 200 個分類器，最後篩選出了 47 個貢獻度較大的變數。對於 $\ell 1$ 正則化，我們發現當學習速率$\boldsymbol{\eta}$=0.0005 時模型預測性能最好，在此條件下篩選出了 50 個變數。

6.2 通過 SMOTEENN 算法進行欠采樣過采樣

欠採樣過採樣技術包括樸素隨機欠採樣演算法、樸素隨機過採樣演算法、SMOTE 演算法、ADASYN 演算法、SMOTEENN 演算法和 SMOTETOMEK 演算法。我們將 5 種類別平衡技術用在三種 Boosting 模型中，結果發現 SMOTENN 演算法對類別的平衡能起到最好的作用，對模型預測我們所關注的類別（即興趣程度 “高” ）的性能的提升最大。過採樣結果如圖 19 所示。

6.3 建模過程

6.3.1 模型效果的评价标准

通過對模型評估指標的一系列對比，考慮到本案例是多分類問題，因此二分類問題中常用的評估指標如 Precision，recall，$\boldsymbol{F_2}$ score 和 AUC 等就不適用於本案例，最終我們選用 Accuracy 作為模型效果的評價指標。

將真實類別與學習器預測類別的組合劃分為真正例(true positive) 、假正例 (false positive) 、真反倒(true negative)、假反例 (false negative) 四種情形，令 TP、FP、TN、FN 分別表示其對應的樣例數，則 Accuracy 可用數學運算式表示為：

$$Accuracy = \frac{TP + TN}{P + N}$$

6.3.2 模型的參數及優化

我們將通過模型預測的總準確率以及興趣程度低、中、高三類的準確率，來確定各個模型中分類器的個數以及每個模型對應的特徵篩選方法。

6.3.2.1 GBDT 模型

使用 ℓ1 正則化（即 Lasso）特徵篩選方法、隨機森林方法得出的數據，訓練 GBDT 模型的結果都顯示，在分類器個數分別取 50, 250, 450, 650, 850 時，預測的總準確率隨著分類器個數的增加而增加，興趣程度低、中、高三類的每一類準確率也隨著分類器個數的增加而增加。因此，各模型中的分類器個數都設定為 850。此時，由於使用隨機森林篩選所得數據總準確率以及各類準確率均高于使用 lasso 方法得出的數據，因此對于 GBDT 模型，我們選用隨機森林作爲特徵篩選的方法，并將模型分類器個數設為 850，結果如圖 20、21 所示。

6.3.2.2 XGBoost 模型

使用 lasso 特徵篩選方法、隨機森林方法得出的數據，訓練 XGBoost 模型的結果都顯示，隨著分類器個數的增加，預測的總準確率以及每一類的準確率有上升趨勢。因此，分類器個數選取 850。而兩種特徵篩選方法得出的結果各有高低，使用 lasso 篩選所得數據總準確率和興趣程度爲中等的預測準確率略高于隨機森林方法所得數據，但興趣程度低和高的兩類資料預測的準確度則略低。由于相較于其他房源，我們更關注能够預測出興趣程度高的房源，因此隨機森林方法（預測興趣程度高的一類準確率更高的）成爲我們的選擇。結果如圖 21、22 所示。

6.3.2.3 Adaboost 模型

同訓練 GBDT 模型與 XGBoost 模型相同，使用 lasso 特徵篩選方法、隨機森林方法得出的數據，訓練 Adaboost 模型得到的四種準確率隨著分類器個數的增加而提高。但是分類器個數從 650 增長到 850 時，準確率上升十分緩慢，趨于平穩。我們認爲這是模型出現了過擬合的問題。因此，分類器個數均選取 650，而不是 850。此時，隨機森林方法得出的所有準確率均高于 lasso 方法。因此我們最終選擇隨機森林方法進行特徵篩選。結果如圖 23、24 所示。

6.3.2.4 模型與篩選變量方法小結

綜上所述，GBDT 模型、XGBoost 模型和 Adaboost 模型都選擇隨機森林方法進行特徵篩選，GBDT 模型、XGBoost 模型分類器個數確定爲 850，Adaboost 模型分類器個數選擇 650。我們將使用這三種組合，預測 test 集，來確定最終的最優模型組合。

6.3.3 模型結果比較

在最後模型的比較中，根據上文的結論，筆者將使用選定的三種 boosting 模型組合對測試集進行預測。我們選取除測試集以外的所有數據作爲訓練集，使用隨機森林方法對訓練集進行特徵篩選，幷進行過采樣欠采樣處理。最終，三種方法的結果爲：GBDT 的總準確率爲 0.6812，高類準確率爲 0.3553；XGBoost 的總準確率爲 0.5413，高類準確率爲 0.1986；Adaboost 的總準確率爲 0.6302，高類的準確率爲 0.3419。（所有類的準確率如圖 26 所示）鑒于 GBDT 模型總準確率與高類的準確率均爲三個模型中的最高，我們最終選擇 GBDT 作爲最終的模型。

6.3.4 對最終模型結果的思考

通過對于 test 集的預測，我們發現房屋租賃興趣程度高的一類準確率大幅降低（相較于 validation 集的預測準確率）。因爲我們十分關注房屋租賃興趣程度高的這一類，所以我們將探尋導致此結果的原因。

首先，數據本身的特殊性可能會導致這一結果。在數據中的 1500 多個變量中，存在著一個由 1500 個變量組成的稀疏矩陣，這個龐大的稀疏矩陣在進行了過采樣欠采樣後，可能会影響測試集高類的預測準確率。因此，我們采用控制變量法，在自變量中刪去這個稀疏矩陣，其他操作不變。儘管如此，對于高類的預測準確率依然只有不足 20%，所以稀疏矩陣影響最終結果的可能性極小。

其次，採樣法也可能導致準確率的降低。因此，我們在略過過采樣欠采樣的步驟的情況下，按照既定方法通過 GBDT 模型進行建模預測，所得結果為：GBDT 的總準確率爲 0.7501，高類準確率爲 0.5770。0.577 遠大于進行採樣法處理後的

高類準確率 0.3553。因此，我們得出結論：在此數據中，採樣法的處理得確導致了高類預測準確率的降低。這是因爲 Boosting 的運作原理與採樣法處理發生了矛盾。Boosting 將一堆弱分類器叠加，提升成一個强分類器。它會根據前一次估計的結果改變後一次的策略，對前一次估計不佳的加大權重，把權重大的那些數估計好一些。所以如果應變量每一類的數量有差距，boosting 會調節權重，處理好這些差距。然而使用過采樣欠采樣將每一類的個數調成一致，反而不利于 boosting 權重的調節與模型的建立，從而降低了預測的準確率。

第七章 結論與展望

基於本文在研究和分析過程中的不足，我們提出以下研究展望：

1. 在變數篩選方式與模型種類相同的情況下，基於採樣後樣本所構建的模型效果劣於基於未經採樣樣本所構建的模型。基於 Boosting 演算法原理，我們做出以下推斷：在 Boosting 多分類問題中，若樣本分佈不均勻，在預測錯誤時，模型會為預測錯誤的樣本增減權重，以此達到優化結果的目的；在經過採樣法後，樣本分佈均勻，因此模型對少數類樣本給出的權重會受到影響，因此導致少數類樣本預測準確率降低，多數類樣本預測準確率上升。因此，我們建議在基於 Boosting 演算法的多分類問題中，慎重考慮是否採用過採樣或欠採樣方法，以此達到提升少數類樣本預測準確率的目的。

2. 本文在評估模型的過程中，所使用的指標為預測準確率。當樣本分佈不均勻時，某一種類樣本比例偏高將很有可能導致評價指標無法準確反映模型優劣。因此我們認為，採用為指標中各項賦予權重的方式將會提升評估效率。例如，通過增加少數類樣本的權重，同時降低多數類樣本的權重，從而達到平衡各類誤差的目的。

3. 本文的有關研究都是基於 Renthop 網站所提供的紐約市公寓出租資料，由於每個城市及網站的目標客戶群體都具有獨特性，在今後的研究中，可將範圍擴大到其他的國家、城市及平臺，從而提高研究的普適性。

4. 由於精力和時間的限制，本文在模型的構建過程中，並沒有選擇其他種類的模型建立完整的對照組，且在變數選擇的過程中步長過大，很有可能導致結論資料為非最優解。因此，在今後的研究中會增加模型數量，縮小參數調整步長，以達到提高說服力的目的。

5. 由於筆者能力有限，無法對資料集中的圖片進行有效處理。在今後的研究中將通過卷積神經網路等方法提取圖片資訊，以增加資料的完整度。

參考文獻

Kain, J. F., & Quigley, J. M. (1972). Housing market discrimination, home-ownership, and savings behavior. *The American Economic Review*, *62*(3), 263-277.

Quigley, J. M. (1976). Housing demand in the short run: An analysis of polytomous choice. In *Explorations in Economic Research, Volume 3, number 1* (pp. 76-102). NBER.

Li, M. M. (1977). A logit model of homeownership. *Econometrica: Journal of the Econometric Society*, 1081-1097.

Silberman, J., Yochum, G., & Ihlanfeldt, K. (1982). Racial differentials in home purchase: The evidence from newly-formed households. *Economic Inquiry*, *20*(3), 443.

Shear, W. B., Wachter, S. M., & Weicher, J. C. (1988). Housing as an Asset in the 1980s and 1990s. *Housing Fin. Rev.*, *7*, 169.

Perloff, J. M. (1991). Choice of housing tenure and wage compensation of hired agricultural workers. *Land Economics*, *67*(2), 203-212.

VanderHart, P. G. (1994). An empirical analysis of the housing decisions of older homeowners. *Real Estate Economics*, *22*(2), 205-233.

Eppli, M., & Childs, M. (1995). A descriptive analysis of US housing demand for the 1990s. *Journal of Real Estate Research*, *10*(1), 69-86.

Bourassa, S. C. (1995). A model of housing tenure choice in Australia. *Journal of Urban Economics*, *37*(2), 161-175.

VanderHart, P. G. (1998). The housing decisions of older households: A dynamic analysis. *Journal of Housing Economics*, *7*(1), 21-48.

Kan, K. (2000). Dynamic modeling of housing tenure choice. *Journal of Urban Economics*, *48*(1), 46-69.

Boehm, T. P., & Schlottmann, A. M. (2004). The dynamics of race, income, and homeownership. *Journal of Urban Economics*, *55*(1), 113-130.

Belsky, E. S., & Belsky, E. S. (2013). *The dream lives on: The future of homeownership in America*. Cambridge, MA: Joint Center for Housing Studies, Harvard University.

Jansson, T. (2017). Housing choices and labor income risk. *Journal of Urban Economics*, *99*, 107-119.

Zhao, J., Chen, L., Pedrycz, W., & Wang, W. (2019). Variational Inference-Based Automatic Relevance Determination Kernel for Embedded Feature Selection of Noisy Industrial Data. *IEEE Transactions on Industrial Electronics (1982)*, *66*(1), 416–428.

Breiman, L. (2001a). “Random forests.” *Machine Learning,* 45(1): 5-32.

Breiman, L., J. Friedman, C. J. Stone, and R.A.Olshen. (1984). Classification and Regression Trees. Chapman & Hall/CRC, Boca Raton, FL.

Quinlan, J.R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Wan L, Zeiler M, Zhang S, et al. (2013). Regularization of neural networks using dropconnect. *International Conference on Machine Learning*, 1058-1066.

Koziarski, M. (2020). Radial-Based Undersampling for imbalanced data classification. *Pattern Recognition*, *102*.

Chawla, N., Hall, L., Bowyer, K., & Kegelmeyer, W. (2002). SMOTE: Synthetic minority over-sampling technique. *The Journal of Artificial Intelligence Research*, *16*, 321–357.

WENG ZHOU, & XIANG LIANG. (2020). *User click prediction method based on gradient boosting decision tree*.

Hastie, T., Rosset, S., Zhu, J., & Zou, H. (2009). Multi-class Adaboost. *Statistics and its Interface*, 2(3), 349-360.

曾珍,邱道持,李鳳,李小廣.大學畢業生租房消費意願及其影響因素研究——基於重慶市的實證研究[J].西南大學學報(自然科學版),2012,34(10):124-130.

陳曉妍. 大學畢業生租房消費意願研究[D].新疆農業大學,2013.

彭秀明. 上海市租賃住房供需研究[D].上海工程技術大學,2013.

路征,楊宇程,趙唯奇.城市外來務工人員租房需求與影響因素分析——基於成都外來務工人員的調查[J].湖南農業大學學報(社會科學版),2016,17(04):89-95+102.

莊靜,李夢微.租售同權背景下南京市即將步入社會大學生租房意願及影響因素實證分析[J].黑龍江科學,2018,9(15):16-17.

孫曉輝,劉寶貞,王婷婷,劉璿.租售同權政策對中低收入者的租房意願影響研究[J].廣西品質監督導報,2019(04):205.

劉靈輝,邱曉豔,王科宇.農村集體建設用地建設租賃性住房的租住意願影響因素研究——基於 1003 名大學生的調研[J].西南交通大學學報(社會科學版),2019,20(06):117-128.

陳珍珠. 基於供給主體的鄭州市住房租賃市場發展路徑研究[D].鄭州大學,2020.

附錄

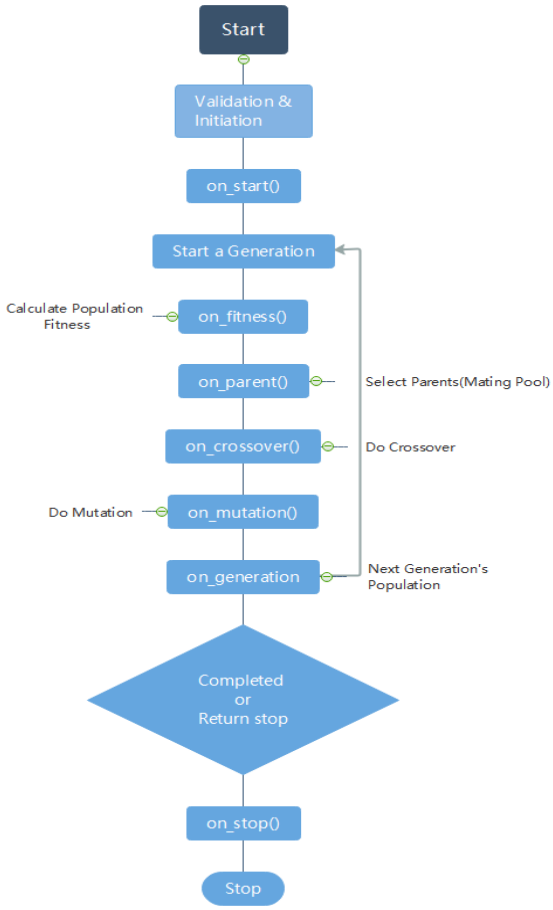

圖 1 遺傳算法流程圖

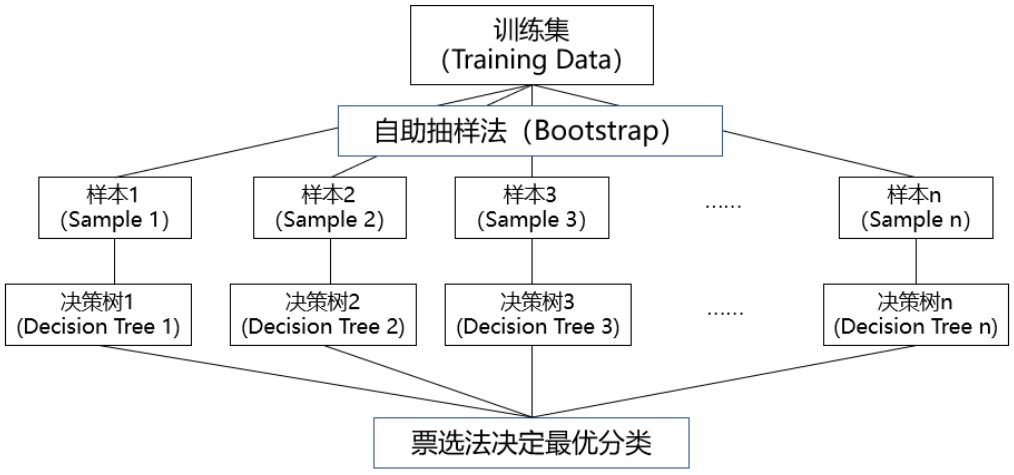

圖 2 隨機森林流程圖

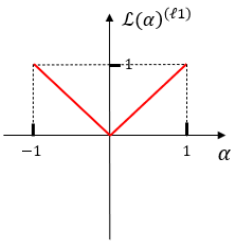

圖3

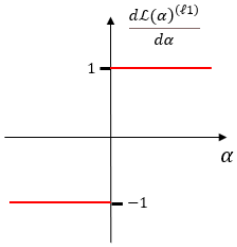

圖4

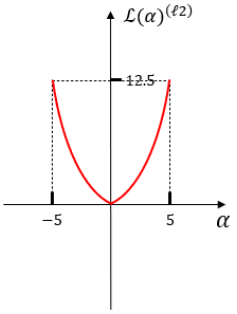

圖5

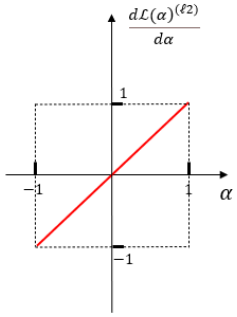

圖6

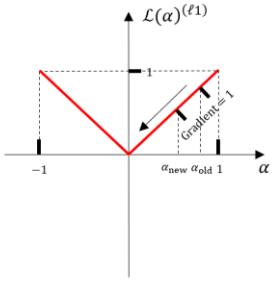

圖7

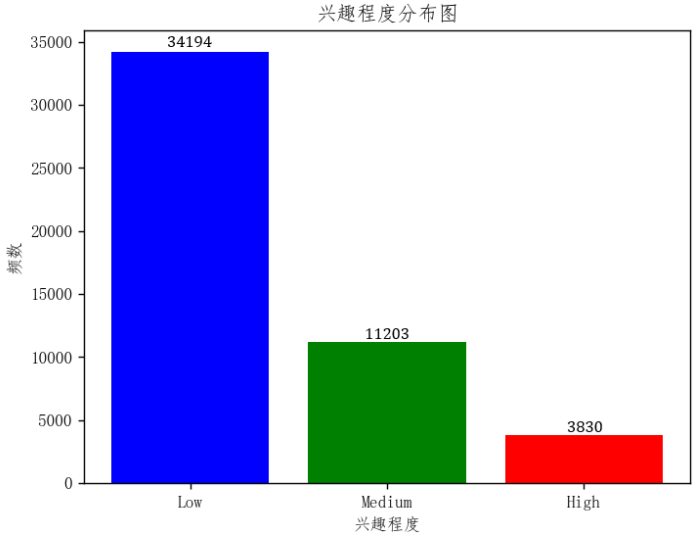

圖 8 興趣程度分佈圖

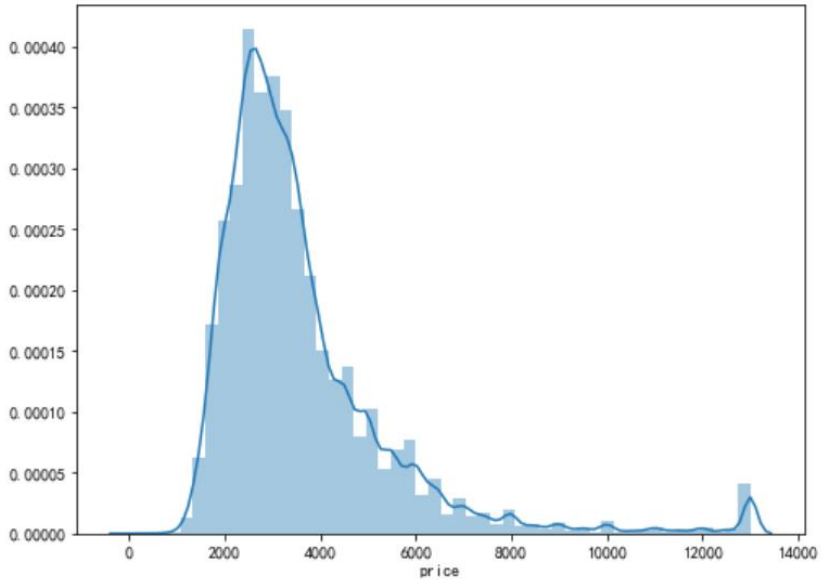
0.00040
0.00035
0.00030
0.00025
0.00020
0.00015
0.00010
0.00005
0.00000
0
2000
4000
6000
8000
10000
12000
14000
price

圖 9

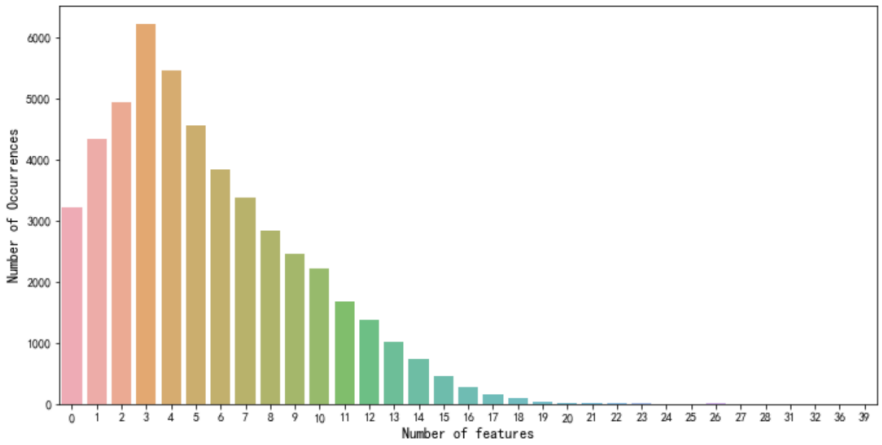
Number of Occurrences
Number of features

圖 10

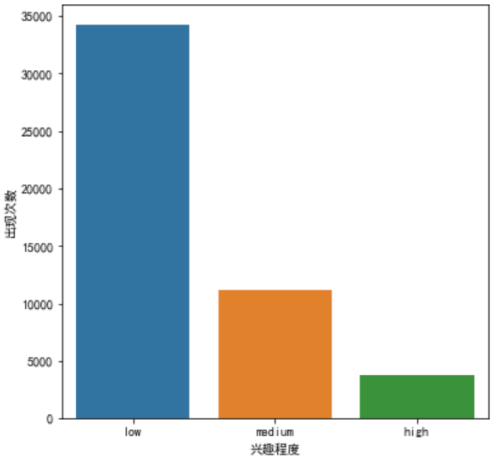
出现次数
low
medium
high
兴趣程度

圖 11

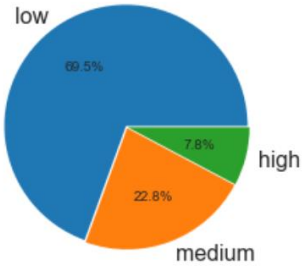

圖 12

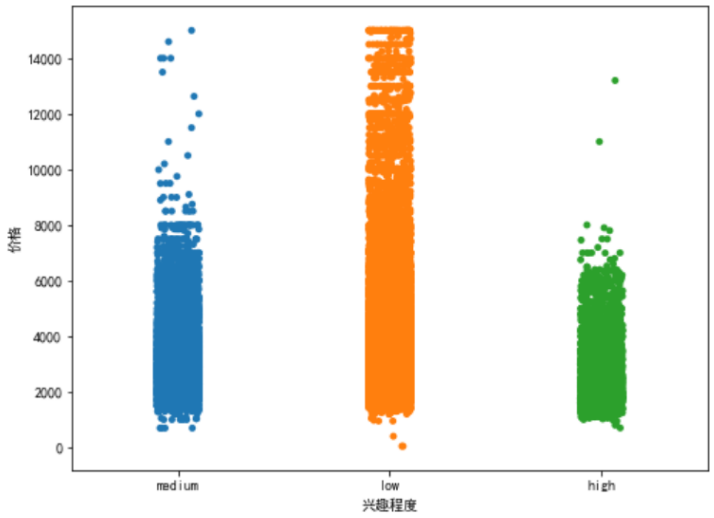

圖 13

```
[Laundry in Unit, Dishwasher, Hardwood Floors, No Fee]
[Elevator, Cats Allowed, Dogs Allowed, Exclusive]
[Common Outdoor Space, Cats Allowed, Dogs Allowed, Doorman, Elevator, Fitness Center, Laundry In Building]
[Doorman, Fitness Center, Dogs Allowed, Cats Allowed]
[Pre-War, No Fee, Dogs Allowed, Cats Allowed]
[Doorman, Elevator, Fitness Center, Cats Allowed, Dogs Allowed]
[Doorman, Elevator, Laundry in Building, Dishwasher, Hardwood Floors]
[Fitness Center, Cats Allowed, Dogs Allowed]
[Elevator, Dishwasher, Hardwood Floors]
[Doorman, Pre-War, Dogs Allowed, Cats Allowed]
[Doorman, Dogs Allowed, Cats Allowed]
[Laundry in Unit, Dishwasher, Hardwood Floors, No Fee, Dogs Allowed, Cats Allowed]
[Elevator, Hardwood Floors]
[prewar, Dogs Allowed, Cats Allowed, LOWRISE, SIMPLEX, HARDWOOD]
[No Fee, Cats Allowed, Dogs Allowed]
[Elevator, Laundry in Building, Dishwasher, Hardwood Floors]
[Dishwasher, Hardwood Floors, Dogs Allowed, Cats Allowed]
[Cats Allowed, Dogs Allowed, Exclusive]
```

圖 14

	bathrooms	bedrooms	building_id	created	description	display_address	features	latitude	listing_id	longitude	
0	1.0	1	8579a0b0d54db803821a35a4a615e97a	2016-06-16 05:55:27	Spacious 1 Bedroom 1 Bathroom in Williamsburgl...	145 Borinquen Place	[Dining Room, Pre-War, Laundry in Building, Di...	40.7108	7170325	-73.9539	a10db4
1	1.0	2	b8e75fc949a6cd8225b455648a951712	2016-06-01 05:44:33	BRAND NEW GUT RENOVATED TRUE 2 BEDROOMFind you...	East 44th	[Doorman, Elevator, Laundry in Building, Dishw...	40.7513	7092344	-73.9722	955db3
2	1.0	2	cd759a988b8f23924b5a2058d5ab2b49	2016-06-14 15:19:59	**FLEX 2 BEDROOM WITH FULL PRESSURIZED WALL**L...	East 56th Street	[Doorman, Elevator, Laundry in Building, Laund...	40.7575	7158877	-73.9625	c8b10
3	1.5	3	53a5b119ba8f7b61d4e010512e0dfc85	2016-06-24 07:54:24	A Brand New 3 Bedroom 1.5 bath ApartmentEnjoy ...	Metropolitan Avenue	[]	40.7145	7211212	-73.9425	5ba98
4	1.0	0	bfb9405149bfff42a92980b594c28234	2016-06-28 03:50:23	Over-sized Studio w abundant closets. Availabl...	East 34th Street	[Doorman, Elevator, Fitness Center, Laundry in...	40.7439	7225292	-73.9743	2c3b4
...	...	...	...	...	...	...	...	...	...	...	...

圖 15

	Unnamed: 0	block	month	index	bathrooms	bedrooms	building_id	created	description	display_address	...	windowed kitchen	c
0	0	Brooklyn	6	4	1.0	1	8579a0b0d54db803821a35a4a615e97a	2016-06-16 05:55:27	Spacious 1 Bedroom 1 Bathroom in Williamsburgl...	145 Borinquen Place	...	0.0	
1	1	Manhattan	6	6	1.0	2	b8e75fc949a6cd8225b455648a951712	2016-06-01 05:44:33	BRAND NEW GUT RENOVATED TRUE 2 BEDROOMFind you...	East 44th	...	0.0	
2	2	Manhattan	6	9	1.0	2	cd759a988b8f23924b5a2058d5ab2b49	2016-06-14 15:19:59	**FLEX 2 BEDROOM WITH FULL PRESSURIZED WALL**L...	East 56th Street	...	0.0	
3	3	Brooklyn	6	10	1.5	3	53a5b119ba8f7b61d4e010512e0dfc85	2016-06-24 07:54:24	A Brand New 3 Bedroom 1.5 bath ApartmentEnjoy ...	Metropolitan Avenue	...	0.0	
4	4	Manhattan	6	15	1.0	0	bfb9405149bfff42a92980b594c28234	2016-06-28 03:50:23	Over-sized Studio w abundant closets. Availabl...	East 34th Street	...	0.0	

圖 16

block	building_id	created	description	display_address
Brooklyn	8579a0b0d54db803821a35a4a615e97a	2016-06-16 05:55:27	Spacious 1 Bedroom 1 Bathroom in Williamsburgl...	145 Borinquen Place
Manhattan	b8e75fc949a6cd8225b455648a951712	2016-06-01 05:44:33	BRAND NEW GUT RENOVATED TRUE 2 BEDROOMFind you...	East 44th
Manhattan	cd759a988b8f23924b5a2058d5ab2b49	2016-06-14 15:19:59	**FLEX 2 BEDROOM WITH FULL PRESSURIZED WALL**L...	East 56th Street
Brooklyn	53a5b119ba8f7b61d4e010512e0dfc85	2016-06-24 07:54:24	A Brand New 3 Bedroom 1.5 bath ApartmentEnjoy ...	Metropolitan Avenue
Manhattan	bfb9405149bfff42a92980b594c28234	2016-06-28 03:50:23	Over-sized Studio w abundant closets. Availabl...	East 34th Street
Manhattan	300d27d8ba2adbcbc8c6f2bcbc1c6f9d	2016-06-28 05:59:06	This spectacular converted 3 bed apartment all...	East 16th Street
Manhattan	0d01cabe55fa5192cdbcabd5c585c1ea	2016-06-08 06:21:36	AMAZING DEAL!! BRAND NEW RENOVATIONS IN THIS H...	East 13th Street
Manhattan	d48767c37a934daaf0bbb0e58c755d0c	2016-06-05 05:28:22	No Fee Large Renovated Sun Splashed Studio. Wa...	York Avenue
Manhattan	d1ca33a2853e64fad6e4009d5d5d168f	2016-06-09 04:42:03	Extra large one bedroom apartment located in P...	E 19 Street
Brooklyn	5f35dc2f0191baf109221752e6ee0c48	2016-06-28 03:26:18	Listed: 06/26/16 Available:...	Hicks Street

圖 17

```
corr_screen = corr[abs(corr)>0.1]
```

```
corr_screen
```

```
interest_label      1.000000
No Fee              0.132051
Hardwood Floors     0.112752
Reduced Fee         0.102692
The Bronx           0.102196
Manhattan          -0.106779
building_code      -0.198449
Name: interest_label, dtype: float64
```

圖 18

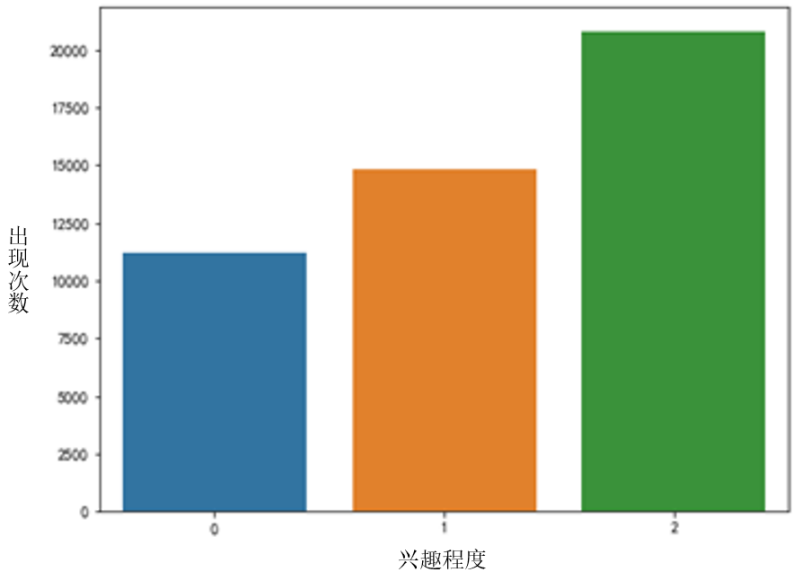

圖 19

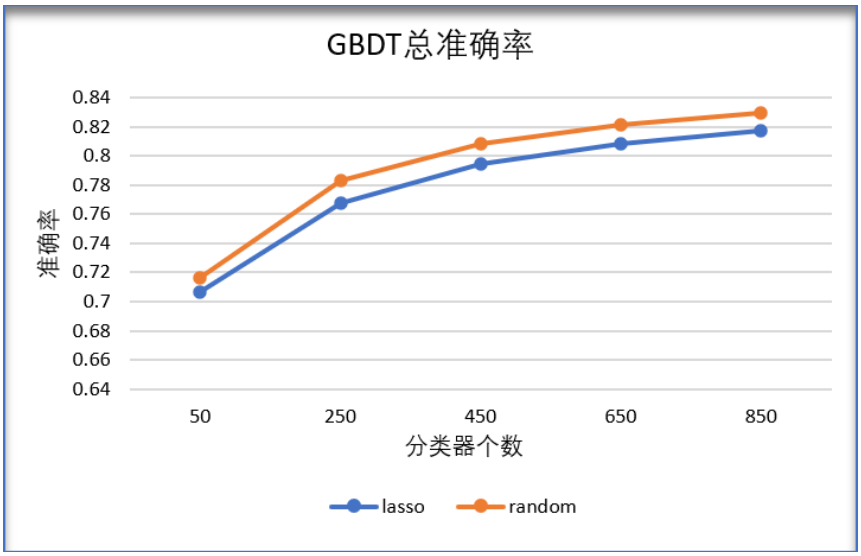

圖 20

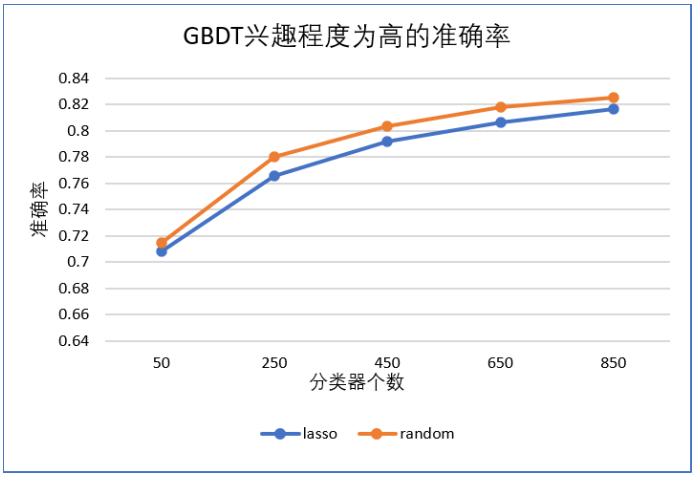
GBDT兴趣程度为高的准确率
0.84
0.82
0.8
0.78
0.76
0.74
0.72
0.7
0.68
0.66
0.64
准确率
50
250
450
650
850
分类器个数
lasso
random

圖 21

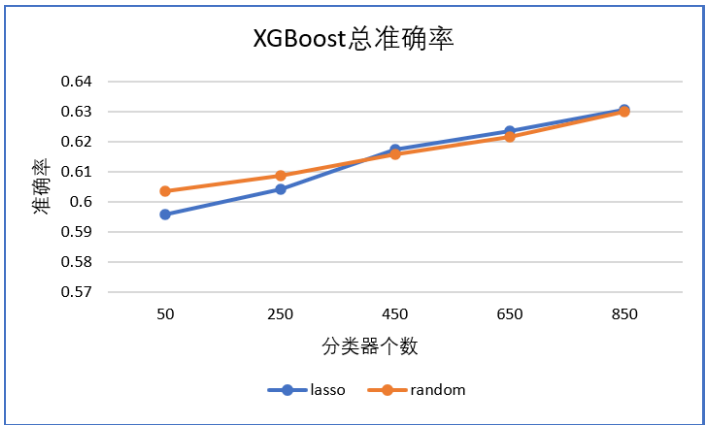
XGBoost总准确率
0.64
0.63
0.62
0.61
0.6
0.59
0.58
0.57
准确率
50
250
450
650
850
分类器个数
lasso
random

圖 22

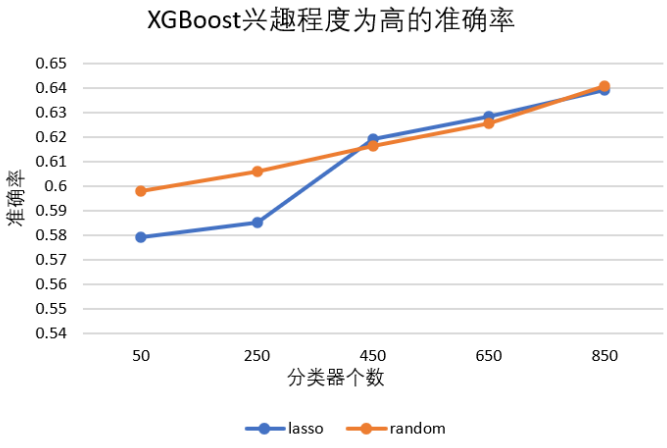
XGBoost兴趣程度为高的准确率
0.65
0.64
0.63
0.62
0.61
0.6
0.59
0.58
0.57
0.56
0.55
0.54
准确率
50
250
450
650
850
分类器个数
lasso
random

圖 23

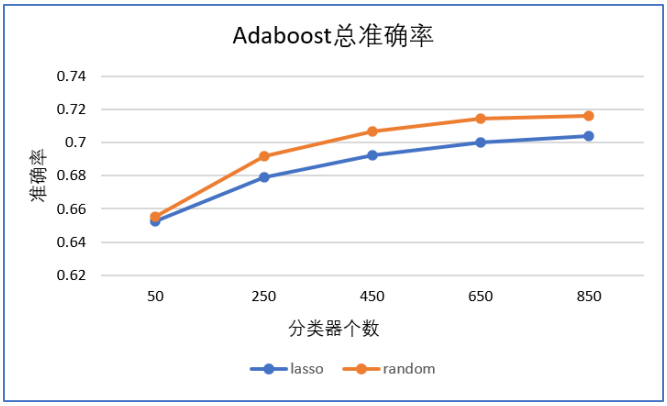

圖 24

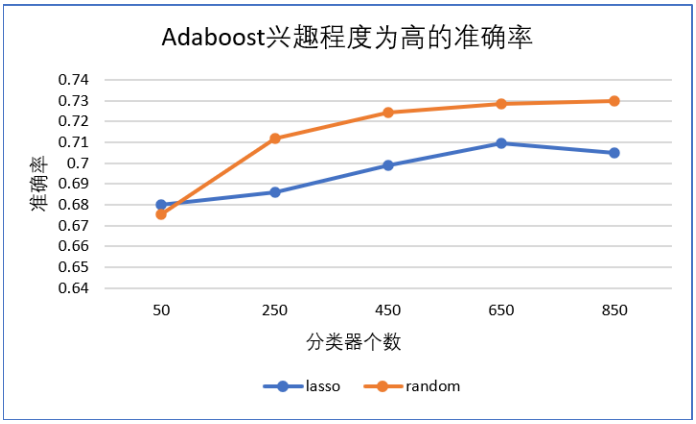

圖 25

	总准确率	低一类的准确率	中一类的准确率	高一类的准确率
GBDT	0.6812	0.8854	0.4091	0.3553
XGBoost	0.5413	0.8834	0.3069	0.1986
Adaboost	0.6302	0.8787	0.3622	0.3419

圖 26

致謝

本次論文話題圍繞商業分析專業課程數據挖掘所學習的內容展開並進行深入探索，基於 Boosting 算法 對租房意願進行預測。本文的完成離不開導師對我的諄諄教導，在接近尾聲之際，我願借此表達誠摯的謝意。

首先感謝我的導師趙鴻昊老師對我的啓發以及教導，雖然建模過程困難重重，趙老師一直展示出非凡的耐心，通宵達旦與我一起解決問題，得以使我的畢業論文順利完成。您不僅激發了我對該領域的興趣，並讓我在該領域積 極地不斷探索創新。在此，謹向尊敬的趙鴻昊教授致以我最真摯的感謝及最熱切的祝福。

我還要感謝工商管理學院的每一位教師，在大學四年通過教授各種課程豐富我們的視野，培養我的批判性思維，令我以一個出色的商學院學生的視角洞察世界，從而逐漸形成屬於自己的人生觀與價值觀。老師們的教導不僅局限于理論知識，更多地是對我獨立思考，獨立學習的能力的訓練，使我成爲一個成熟的個體，一個擁有自我判斷力並擁有美好品德的畢業生。

除此以外，我還要感謝我在大學相遇相知相熟的朋友門，大家的鼓勵與支持在論文完成過程當中發揮著非常積極的促進作用。在交流之中，我們互相學習，擦出思維的火花，爲我的論文撰寫帶來許多靈感與啓發。還要感謝我的家人，在無數個熬夜的夜晚給我們帶來殷切的關懷，令我們倍感溫暖。

“吾生有涯而知無涯”，最後我想感謝澳門科技大學給予我的一個廣闊的平臺，令我在知識的海洋當中不斷專研，結實志同道合之友不斷向前。校方爲我們的成長鋪下了堅實的基石，爲我們創造了許多機會得以歷練我們的綜合能力。 我對此不勝感激。即使畢業季即將來臨，我亦會帶著最澎湃的朝氣迎接未來的機遇亦或者是挑戰，激發自己的潛能，使自己成爲一個優秀善良的人。