Manufacturing Data Science（112-1）
Airbnb Taipei 訂價預測模型與策略優化
Final Project Report

Group G
B10303105 陳郁婷
B10303114 楊喆東
R12725049 徐尚淵
R12725048 楊喻妃

1 Background and Motivation

1.1 Motivation

近年來隨著共享經濟不斷的蓬勃發展，線上短租平台 Airbnb 在全世界掀起一股潮流，更在 2013 年進軍臺灣，全臺的短租型套房也逐漸增加，根據牛津經濟研究院的報告《Airbnb 對台灣之經濟影響》指出 Airbnb 已躋身台灣旅遊業的重要支柱，持續促進經濟成長並創造商機，Airbnb 不單造福旅遊業，旅客在當地的消費活動亦為在地經濟帶來正面影響。隨著越來越多的房東加入這個平台，市場競爭日益激烈，在這樣的環境下，房源的價格擬定變得十分重要，有效的定價不僅影響房東的收入，也直接關係到消費者的選擇和滿意度。

1.2 Background

對住宿業來說，房源的定價直接決定了收益的多寡，過高或過低的價格都可能導致房東和租客雙方權益受損，但是由於共享經濟的房源大多來自於閒置資源，過去對旅館業所使用的訂價參數無法使用，故本專案欲開發可靠的價格預測模型，以幫助租房業者和租客都能夠進行房源的價格預測，以保障租客與房東權益不受損，共同推動健康住宿市場的健康發展。

1.3 Problem Definition

在本專案中，我們採用隨機森林回歸（Random Forest Regressor）模型，從「Inside Airbnb」網站抓取數據，深入分析和識別影響 Airbnb 房源出租價格的關鍵因素。而此模型綜合考慮了房源的類型、佈局、提供的設施與服務等多維度特徵，該模型旨在準確推估市場的支付意願價位，為房東提供更科學、精準的價格建議，幫助他們在競爭

激烈的市場中做出更明智的定價決策。

2 Methodology

2.1 Data Preparation

資料前處理主要分為四個步驟，包含改變資料型態、移除資料、缺失值填補、以及類別變數編碼，以下將依序說明細節。

首先，在改變資料型態的階段，先清理含有雜訊如%、$之欄位並轉換為浮點數，再將含時間戳記之欄位轉為與 2023/09/25 的距離天數，其中 2023/09/25 為此資料集的獲取日期，最後原資料集將是與否存為 T 與 F 值，我們將其分別轉為整數 1 與 0。

接著，移除資料時，由於原資料欄位高達 80 項，因此我們先利用人工判斷的方式，將與房價較不相關之欄位移除，其中包含 scrape_id（資料爬蟲編號）、host_name（房東姓名）等。除此之外，由於本次專案目標為預測房價，因此我們以大於第一四分位數加 1.5 倍 IQR 距離，和小於第三四分位數減 1.5 倍 IQR 距離的點為標準，將房價之離群值移除，以避免影響預測結果。最後，針對兩兩高度相關之特徵，我們只保留其中一項，以避免共線性。

再者，關於缺失值，因類別變數缺失值相對較少，在上千筆數據當中只有約 30 個缺失值，因此我們利用眾數填補類別變數，並利用 MICE 填補數值變數，用以減少填入平均值所產生的偏差以及資訊量的流失。

最後，關於類別變數編碼，原資料集中 bathroom_text 欄位儲存各房子的浴室樣態，如 6.5 shared baths、1 private bath 等，由於類別眾多且帶有複合資訊（數量 + 種類），因此將 shared bath、private bath 等作為獨立的欄位，分別儲存相對應之數量。除此之外，原資料集中 amenities 欄位以文字述敘各房子的設備，我們將原始文字斷詞，並清洗標點符號、空格等雜訊，而由於設備種類過多，因此利用人工將其分成較大的類別，如廚房用品、嬰兒用品等，再進行獨熱編碼。最後，針對種類較為單純之類別變數如 room_type、property_type 等，直接予以獨熱編碼。

2.2 Feature Selection

特徵篩選的部份我們採用了兩種方法來進行比較，分別為 Elastic Net 以及投票法（Ensemble Voting），以下便會針對上述兩種方法來進行實作方法介紹。

首先 Elastic Net 同時包含 Lasso 和 Ridge。這意味其既可以可以選擇特徵也能夠達到縮小係數的效果，並處理多重共線性問題，使得所篩選出來的特徵都能夠對房價做出有效的預測效果。

利用 Elastic Net 方法我們一共篩選掉了 20 個不重要的變數，變數由原本的 184

個降為 164 個，以下為我們利用 Elastic Net 方法篩選出的前幾重要的變數，若數值為正則代表其對於房價有正面的影響，數值增加一能夠增加多少房價的價格，而若係數為負則反之。

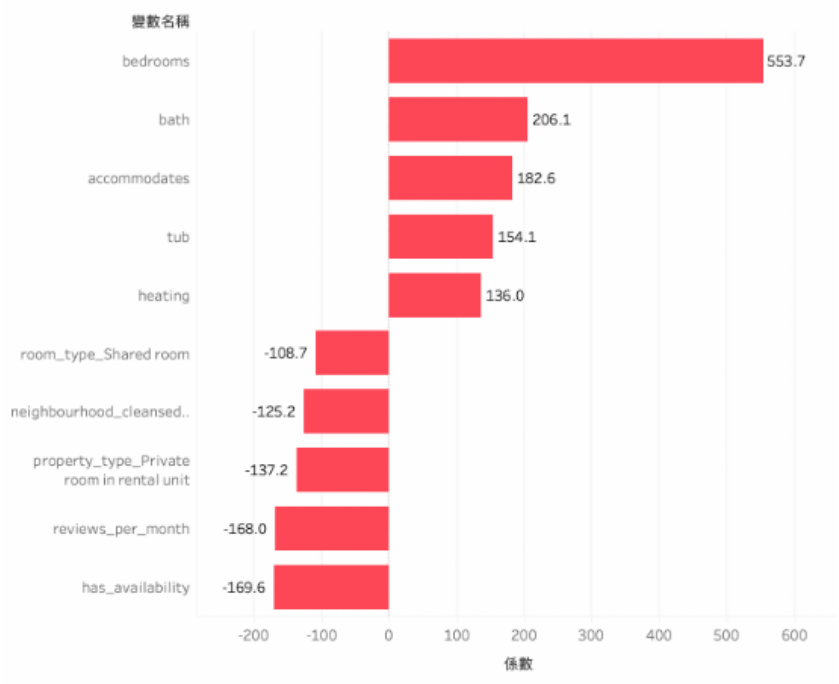

Figure 1: Elastic Net 重要係數排序

而投票法的部分則採用了四種不同的方法，其分別為 RandomForest、XGBoost、Lasso 以及 Stepwise，我們期望此種方法能夠減少單一模型的偏誤，提高特徵選擇的穩健程度並適應多種不同類的數據結構，讓我們同時考量線性與非線性關係和特徵之間的交互作用。

投票法實作後，我們將得到四票的 34 個特徵分為一組訓練資料，而得到四票以及得到三票的 94 個特徵分為另一組訓練資料，加上原有的 Elastic Net 方法便得到三種不同數量特徵的資料集，以供後續模型預測比較使用。

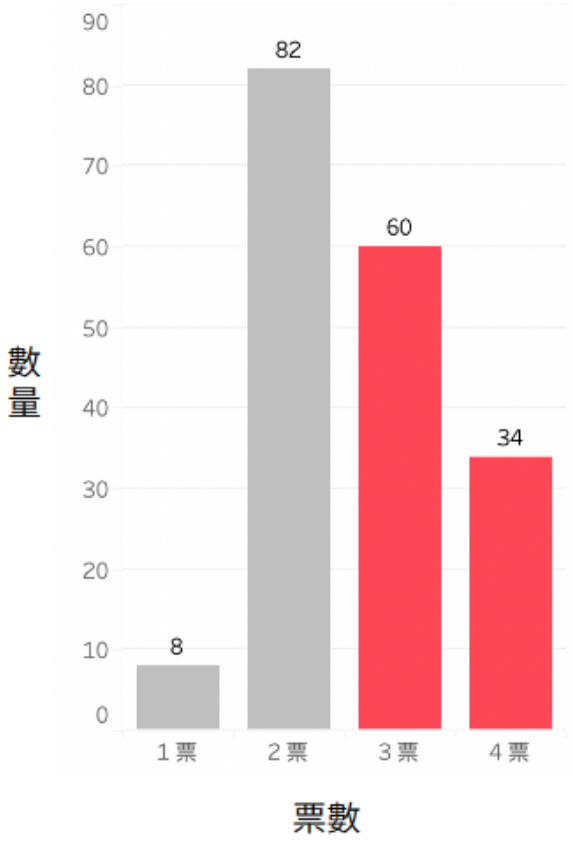

Figure 2: 投票法特徵票數分布

2.3 Prediction Model

建立模型做資料預測的部分，我們選用了四種不同的模型分別為 Linear Regression、XGBoost、SVR、Random Forest，對三個不同數量特徵的資料集做訓練以及預測。對其模型建立的流程也有建立統一的標準，避免因其他應為控制因素的造成模型的比對不夠準確，以下為模型訓練的流程：

1. 切分資料集：將資料集比例統一為訓練資料 0.64，驗證資料 0.16 以及測試資料 0.2。
2. 模型訓練：針對各模型分別以 GridSearchCV 方法來優化超參數，並以 5-Fold Cross Validation 避免 Overfitting。
3. 結果比較：以 RMSE 與 Adjusted R squared 兩個數值指標，比較各模型之預測結果，選出效果最佳者作為本次專案的最佳解。

透過以上流程我們得出根據三種不同資料集訓練出的 RMSE 與 Adjusted R squared 結果數值比較。由下表可看出以 Elastic Net 所挑選出之重要變數建立模型，

其中效能最好的為 Random Forest Regressor、其次則為 SVR。

	RMSE	Adjusted R-Squared
Random Forest	1017.306	0.650
Linear	1255.223	0.384
SVR	1267.424	0.424
XGBoost	1275.348	0.366

Table 1: 以 Elastic Net 所挑選出之重要變數建立模型

由投票法所選出之模型建立變數，同樣是 Random Forest Regressor 的效能以及數值最為理想，最高的 Adjusted R-Squared 值可達 0.681，而其次則為 XGBoost 而非 SVR，此處是與以 Elastic Net 所挑選出之重要變數建立模型的不同之處。

	RMSE	Adjusted R-Squared
Random Forest	988.679	0.681
Linear	1263.997	0.427
SVR	1292.415	0.351
XGBoost	1276.693	0.444

Table 2: 以投票法得四票所挑選出之重要變數建立模型

	RMSE	Adjusted R-Squared
Random Forest	997.085	0.671
Linear	1350.917	0.389
SVR	1281.685	0.411
XGBoost	1276.291	0.413

Table 3: 以投票法得三票以上所挑選出之重要變數建立模型

根據以上三種不同特徵挑選的數據集對應四種不同的模型訓練方式，可以得出 Random Forest Regressor 的預測效果明顯優於另外三者，因此特別將 Random Forest Regressor 獨立出來比較其與三種不同特徵挑選的數據集之間不同的效果，可得出 4x3 的 12 組之中，以投票法四票對應 Random Forest Regressor 效果最佳，因此後續都將以此組來做分析。

	RMSE	Adjusted R-Squared
投票法 4 票	988.679	0.681
投票法 3 與 4 票	997.085	0.671
Elastic Net	1017.306	0.650

Table 4: 以投票法得三票以上所挑選出之重要變數建立模型

3 Data Collection and Analysis Result

3.1 Data Collection

本次的專案，我們取用了 Inside Airbnb 上 2023/9/25 的數據做為研究資料。此網站收集了来自 Airbnb 網站的數據，如出租房子的詳細資訊、定價和其他相關資訊。其目標是使平台增加透明度，幫助用戶更好地了解 Airbnb 對當地住房市場的資訊。

3.2 Analysis Result

我們最終使用的模型為 Random Forest Regressor，且特徵集合是由投票法中得到四票的重要變數組成。下圖為模型所跑出來的結果：

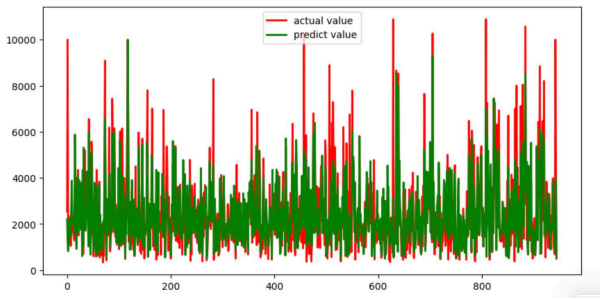

Figure 3: 預測值和實際值

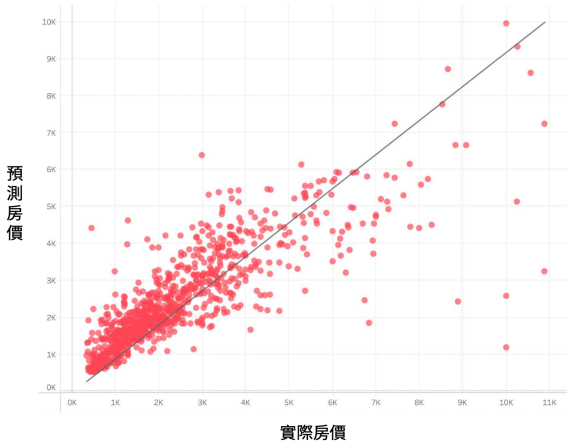

Figure 4: 預測值與實際值散佈圖

接著找出表現最佳的模型，並將其變數依據重要性排名（重要性是指通過考慮每個決策樹中節點分裂所帶來的信息增益的增加量來計算的。當一個特徵在節點上被使用進行分裂時，該特徵的重要性就會增加。最後綜合所有樹的結果，得出每個特徵的相對重要性）：

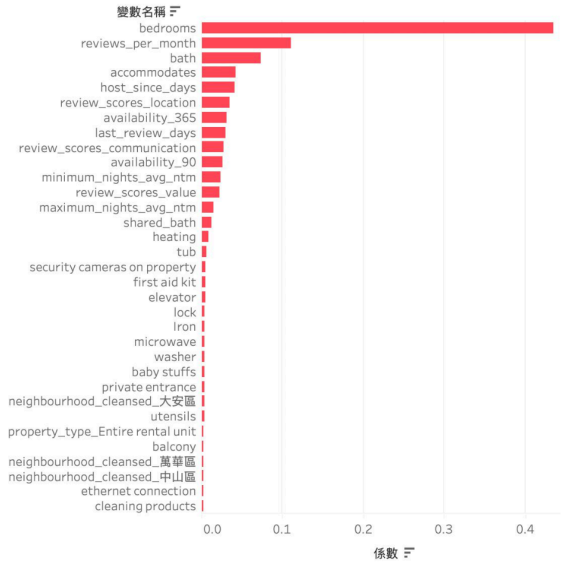

Figure 5: 重要變數排名

可將得出的重要變數大致整理成六大類：

1. 基礎設備：包含清潔、衛浴以及嬰兒用品等。
2. 地點：大安區、中山區、萬華區等區域。

3. 出租狀況：包含租借天數上下限等。

4. 評論：評論數、分數、最新評論日期等。

5. 屋主狀況：屋主年資。

6. 房子型態：是否為包棟建築。

3.3 Interpretation

Figure 6: 使用者可以根據屋源狀態來輸入資訊 (i.e. 地區、屋源類型、房間種類)

Figure 7: 接者輸入更詳細的資訊 (i.e. 可容納人數、床位等等)

Figure 8: 使用者可點選其屋原有提供的設備或服務

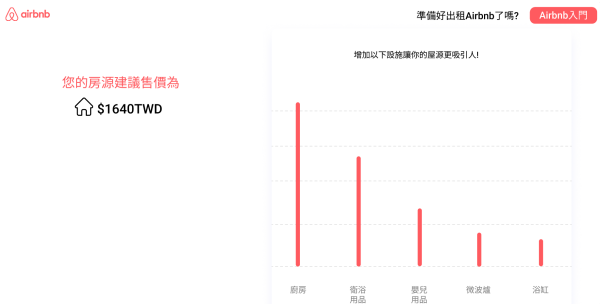

Figure 9: 最後結果畫面，提供使用者適當的願付價格，以及根據重要特徵排序，推薦房東可新增的設備以提高此屋源的願付價格

4 Conclusion

4.1 Conclusion and Suggestion

本專案採取 Inside Airbnb 2023/9/25 的數據，探討影響「Airbnb」房源出租價格之因素，並了解其相關聯性，故在本章節將以最後模型呈現的結果配合 Airbnb 其經營策略，為房東提出行銷策略上的建議，也幫助 Airbnb 改善其平台經營方式。

考慮到房源的格局和類型等特徵在初始階段已固定，且對於房東來說後期調整的難度較大，本專案因此著重於探討房東可在設備與服務層面上進行的優化調整，根據我們的研究結果分析，我們識別出影響租客願付價格的三個關鍵特徵：清潔、衛浴設施和嬰兒用品。這突顯了租客在選擇住宿時對於乾淨、整潔環境的重視，因為直接影響了他們對房源的第一印象，因此乾淨程度和有無衛浴設施，是提高房源吸引力的關鍵要素。

4.2 Future Work

本分析結果雖然呈現了決定一個房屋願付價格的重要特徵，但仍有許多部分受到限制，因此在最後的章節說明本分析受到的限制，並給予未來可深入研究的方向。

1. 資料時間範圍的限制：本研究所使用的數據來自「Inside Airbnb」，其提供的資料主要以季度為單位更新，因此，我們所獲得的最新數據為 2023 年 9 月 25 日的訂價，此更新方式限制了我們深入探究特定日期，如平日、假日或節慶期間，對房價的具體影響，這意味著季節性和特定日期對價格變動的影響可能未能在本分析中得到充分考量。

2. 變數與價格關係的複雜性：根據課堂上教授的反饋，我們意識到並非所有變數都與房價呈線性關係，例如，我們最初假設廁所數量可能是影響願付價格的一個重要特徵，但在實際分析中發現，房間的廁所數量與願付價格之間的關係並非直接

線性升高，這表明我們在未來的研究中需要更仔細地探究各變數與目標變數之間的關聯性，並考慮設定某些變數的上限值，如廁所數量的合理範圍，以更準確地反映其對價格的影響。