House Prices - Advanced Regression Techniques project competition

劉翊安 不分系112 F64081020

1. 專題簡介

本專題基於資料科學專案競賽平台“kaggle”，其中一個名為“House Prices - Advanced Regression Techniques”的比賽。

此專案使用美國一個城市Ames, Iowa的真實房價資料。目標是透過現有的資料預測其他房屋的真實價錢，平台上會有預測精準度的排名，學期一開始的目標是進入PR75以上。

2. 動機＆影響

動機：我將本次專題定義成一個學習型專題，在完成專題後可以達到以下以個目的。

1. 了解資料科學專案的完整流程並學習在一個專案中需要用到的各種不同工具。
2. 變成自己的履歷，作為學涯、職涯中向人毛遂自薦的作品。
3. 把學到的技巧運用在下一個學期的專題中，解決日常生活中實際會遇到的問題。

產出結果：以下產出將會同時放在Kaggle的討論區上。

1. 一個可以良好預測房價的模型。
2. 所有實驗方法及結果的紀錄。
3. 個人對實驗結果的觀察及洞見。

預期影響

1. 提升房價預測準確度：透過建立一個可靠且準確的房價預測模型，我們能夠幫助人們更好地了解和預測房屋價格。這對於房地產行業、投資者和一般民眾都具有重要意義，可以協助他們做出更明智的決策。
2. 促進居住正義：房價是一個社會經濟議題，不同社會階層的人對於房屋的價格承受能力不同。透過精確的房價預測模型，我們可以幫助政府、非營利組織和社區開發者更好地規劃和分配房屋資源，提供更公平和平等的居住機會。
3. 推動資料科學社群交流與學習：透過在Kaggle平台上分享專案的方法和結果，我期望能夠與其他資料科學家進行交流和討論，互相學習和提升。我將分享我的觀察、洞見和實驗結果，以促進知識共享和技術進步。
4. 作為履歷和職涯發展的資產：這個專題將成為我個人的履歷資產，展示我的能力和對資料科學的熱情。我期望通過這個專題的完成，能夠增強我的職業競爭力，為未來的職涯發展打下基礎。

3. 結果－實驗結果、排名進步幅度

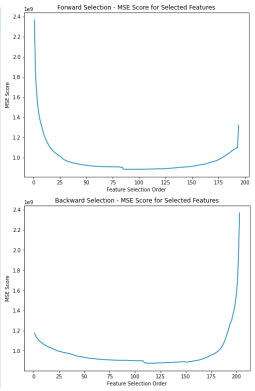
在這個學期中，我一共做了十幾次的實驗，以下會列出我做過的嘗試及結果。

- **簡單的填補缺失值＋XGBoost**：一開始我用簡單的方法填補缺失值，搭配XGBoost訓練模型。可以看到成效不太好，出來的成果排名在Kaggle上只有PR47而已。
- **改良填補缺失值方法**：為了提升預測精準度，我開始對資料的特性做完整的研究。因此可以用更具有邏輯的法填補缺失值，可以看到預測精準度大幅提升，我的排名也因此上升到PR67。
- **改進XGBoost選擇參數的方法**：利用RandomizeCV，總共訓練模型50次，每一次的參數選擇都是隨機從我指定的幾個值裡面取的，做完以後取出結果最好的那一組作為欲使用的參數。預測精準度也因此變得更好一些。
- **特徵選擇(feature selection)**：我嘗試幾種不同的方法Lasso, Random forest, Correlation, Mutual info, forward selection, backward selection，選擇出欲使用的特徵後再做訓練。
- **資料預處理**：將原始資料做standardize, log transformation後再做訓練。

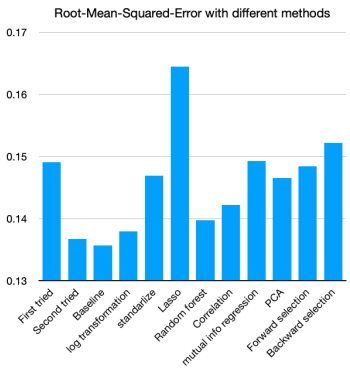
最好的訓練結果出現在使用較好的資料預處理方法加上RandomizeCV以後。雖然接下來的嘗試沒有得到更好的結果，但也因此讓我在過程中學到各種處理資料科學專案的方法，相信對接下來的專題會有很大的幫助！

	Missing value	One hot encoding	Data processing	Feature selection	ML algorithm	result	ranking(pr)
First tried	M0	V	X	X	XGBoost	0.14906	0.47
Second tried	**M1**	V	X	X	XGBoost	0.13672	0.67
Baseline	m1	V	X	X	**XGBoost+randomize CV**	0.13573	0.68
log transformation	m1	V	**Log transformation**	X	XGBoost	0.13791	
standarlize	m1	V	**standarlize**	X	XGBoost	0.14687	
Lasso	m1	V	X	**Lasso**	XGBoost	0.16443	
Random forest	m1	V	X	**Random forest**	XGBoost	0.1397	
Correlation	m1	V	X	**Correlation**	XGBoost	0.14215	
mutual info regression	m1	V	X	**mutual info regression**	XGBoost	0.14925	
PCA	m1	V	X	**PCA**	XGBoost	0.14652	
Forward selection	m1	V	X	**Forward selection**	XGBoost	0.14843	
Backward selection	m1	V	X	**Backward selection**	XGBoost	0.1522	

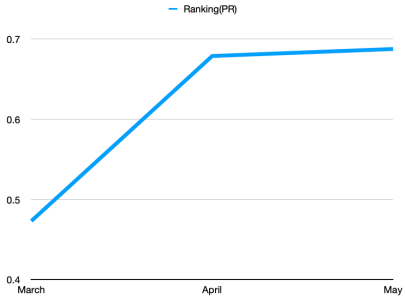
實驗方法及對應的誤差、排名



FORWARD/ BACKWARD SELECTION 特徵數量對應誤差圖



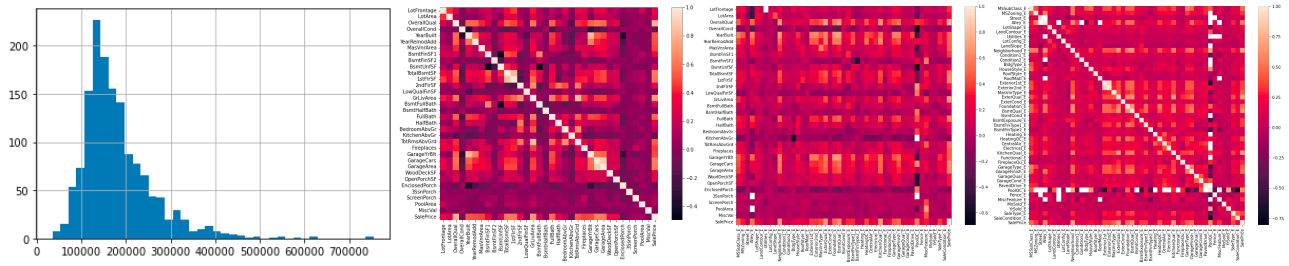
實驗結果(MSE)柱狀圖



排名進步折線圖

4. 方法

1. **Data research**：將房價資料畫成柱狀圖觀察分佈情況；對每一個特徵做Heatmap觀察特徵之間的相關狀況。



2. **Missing value**：共使用兩種不同方法處理缺失值。第二種方法是參考Data research的結果做的，相對較有邏輯。

	Filling missing value	Drop
Method 1	Int -> Mean Object -> 0	More than 70% 5 features
Method 2	Decide with observation ex: Int -> categorical ex: fireplace -> FireplaceQu	More than 75% 4 features

3. **One hot encoding**：將資料型態是類別的特徵用One hot encoding的方法轉為數值形式。

MSZoning
RH
RM
RM
RL
RL

FV	RH	RL	RM
0	1	0	0
0	0	0	1
0	0	0	1
0	0	1	0
0	0	1	0

```
RL          1151
RM           218
FV            65
RH            16
C (all)       10
Name: MSZoning, dtype: int64
```

4. **XGBoost(eXtreme Gradient Boosting)**：XGBoost是我在這一次專題主要使用的演算法。XGBoost通過逐步構建多棵決策樹使結果漸漸逼近實際值，以下是一個簡單的說明圖片。

5. 結論

經過這一次的專題，我對一個資料科的專案該如何進行有更深入的認識。透過實際執行讓我對“如何處理缺失值”、“如何觀察資料並了解資料特性”、“機器學習演算法-XGBoost”、“特徵選擇的各種方法”這些議題有更加了解，同時也花了不少時間深入了解當中的數學模型。
我將會把“良好預測房價的模型”、“實驗方法及結果紀錄”、“個人對實驗結果的觀察及洞見”放在Kaggle的討論區上。期望透過我的分享可以幫助到其他在做相關研究的人，同時與其他資料科學家交流，精進自己的能力。
此次的專題結果可以成為我學涯、職涯中向人介紹的作品，同時也能把學到的技巧運用在下一個學期的專題中，解決日常生活中實際會遇到的問題。
若有機會可以更加精進此次專題的成果，相信這會是一個能夠弭平消費者對房價資訊的落差，落實「居住正義」的好工具！