國 立 臺 北 科 技 大 學

NATIONAL TAIPEI UNIVERSITY OF TECHNOLOGY

課程名稱: 機器學習

實習單元名稱: House Sale Price Prediction Challenge

班級: 電四乙

姓名及學號:劉昆琳 107310203

中　華　民　國110年11月17日

1. 作法說明

此次實驗參考「Stacking」的作法，其是整合多個回歸模型的整合學習技術，模型架構分做兩個部分，第一為**基礎模型**，第二為**元模型**，其中基礎模型利用整個訓練集做訓練，而元模型則將基礎模型的輸出結果作為特徵進行訓練；在本實驗中，基礎模型採用DNN、KNN、及LightGBM，最後利用3層全連階層(Fully Connected Layer)整合三個基礎模型的輸出。

深度神經網路(DNN) : 模擬人類神經網絡的運作方式，可以將其拆為輸入層(Input layer)、隱藏層(Hidden layer)及輸出層(Output layer)，網路中的每一層都有許多神經元(Neuron)，而各層之間神經元與神經元的連接都是倚賴權重(Weight)和偏差(Bias)，權重可以使用梯度下降（Gradient descent）更新。

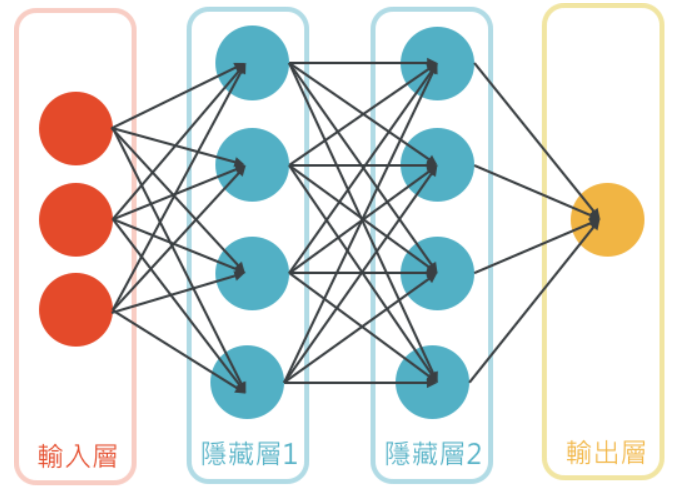


圖1 深度神經網路(DNN)

K-近鄰算法(KNN) : KNN 最近鄰居法又譯為K-近鄰算法，是所有機器學習算法中最簡單，也是使用最廣的演算法之一，其數學原理為透過K個距離最近的鄰居，並依照這些鄰居的房價決定賦予待測點的預測值；以下是KNN每個步驟的說明:

1. 利用歐基里德距離計算待測點與已知每個點之間的距離。
2. 計算完距離以後，需要決定常數K，也就是有多少個鄰居被視為最接近的鄰居。舉例來說，如果選擇K=3，那就會以距離最接近的3個點當作鄰居，並將這3個鄰居的房價取總和後平均，得到的值即為待測點的預測值。

LightGBM : 其是基於樹的學習算法，所謂的決策樹是透過一連串的決策，分出不同結果所組成的樹狀圖，而隨機森林(Random Forest)則是以決策樹為基礎，透過多個隨機決策樹找到合適的答案，並將多個好結果的決策樹聚合成好的隨機森林；隨機森林樹的生長方式是水平方向的，而LightGBM是垂直方向的，也就是說LightGBM生長的是樹的葉子，其他的算法生長的是樹的層次，這樣的好處是速度很快，可以處理大量的數據。

一張含有 文字, 美工圖案 的圖片

自動產生的描述

圖2 LightGBM樹生長方式

1. 程式方塊圖與寫法

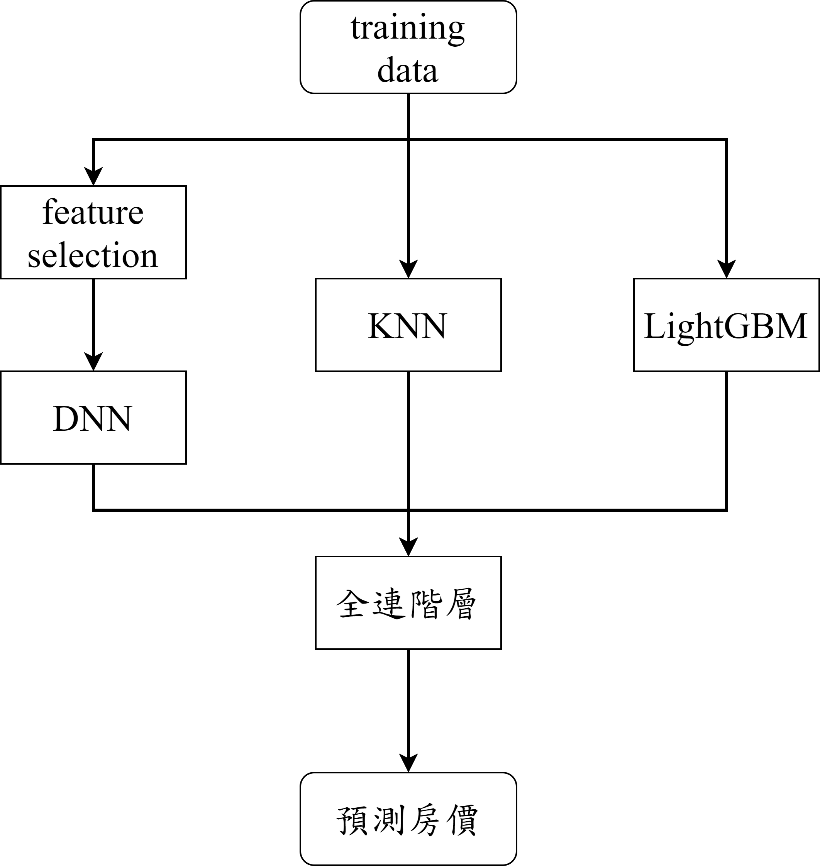


圖3 實驗架構圖

feature selection : 由於提供的22種特徵，有些對於房價並無直接的關聯，如year、month、day，所以須將這些特徵移除，如果沒有移除的話，可能會使模型預測錯誤，故使用Pandas套件中的dataframe.corr()進行分析，並取出與房價相關性高的前11種特徵。

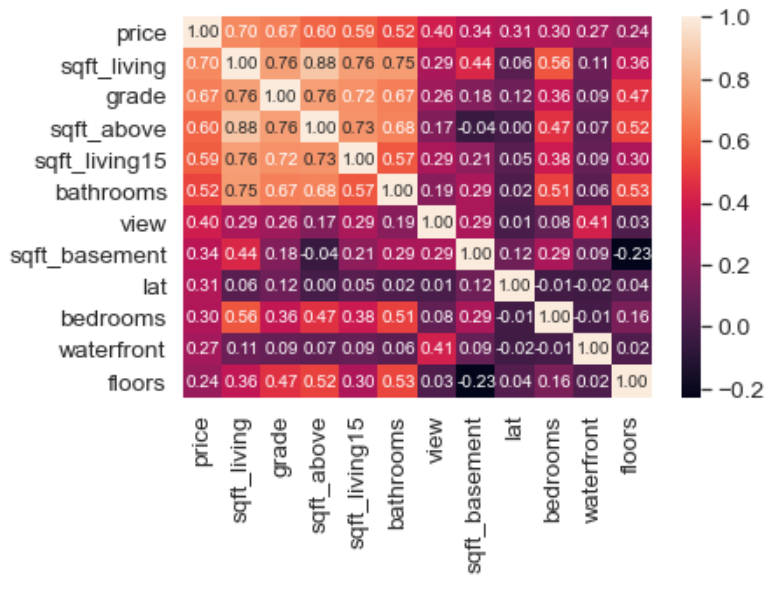


圖4與房價相關性高的前11種特徵

深度神經網路(DNN) : 輸入層為feature selection後的11種特徵，並架設5層的隱藏層，神經元數分別是64、128、512、128、64，激勵函數(Activation function)為relu，此函數能將輸出轉為非線性方程式，公式為，最後由於輸出為房價預測，是一個線性的值，故輸出層的激勵函數將設為linear。

一張含有 桌 的圖片

自動產生的描述

圖5 DNN網路模型

K-近鄰算法(KNN) : 本實驗經由多次的測試後，找到最佳值得K為13，故將待測點距離最近的13個房價總和後平均，即為待測點的預測值。

LightGBM : 採用lightgbm套件匯入，參數設計如圖6，boosting\_type為指定學習器的類型，而設定為gbdt則是使用梯度提升樹；objective為學習目標，由於此次實驗是回歸任務，故設為regression；metric是評估指標，設定為l1、l2分別是用MAE及MSE來評估；num\_leaves為葉子的各數，設定為31；learning\_rate則是學習率，設定為0.05；feature\_fraction為隨機對特徵採樣的比率，設定為0.9；bagging\_fraction為在不進行重新採樣的情況下，隨機選擇部分數據，可以用來加速訓練及處理overfitting，設定之值為0.8；bagging\_freq即每 k 次迭代執行bagging算法，設定之值為5；最後在訓練時則採用early\_stopping的技術防止Overfitting。

一張含有 文字 的圖片

自動產生的描述

圖6 LightGBM參數設定

全連階層整合上述三模型 : 將上述的輸出預測值作為整合模型的輸入層，並架設3層全連接層整合，神經元數分別為8、32、8，且激勵函數為relu，最後輸出層則是房價預測的最終結果，激勵函數將設為linear。

一張含有 桌 的圖片

自動產生的描述

圖7 整合模型

1. 畫圖做結果分析

深度神經網路(DNN) : 由輸出曲線圖可以看到，loss及val\_loss在最後的變化會變得平緩，且無loss下降，val\_loss上升的趨勢，故無Overfitting。

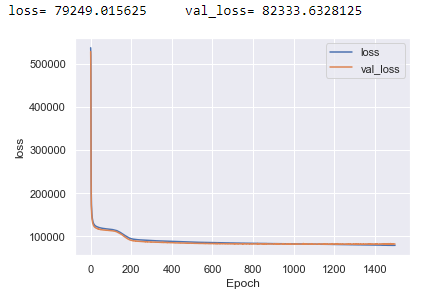


圖8 DNN的loss曲線圖

K-近鄰算法(KNN) : 求得最佳值得K為13，可以利用最近的13個鄰居算出待測點之值，並算出loss為85526.24710782041。

LightGBM : 由圖9可得知，LightGBM的loss有在持續的下降，代表模型有學習到有用的特徵。

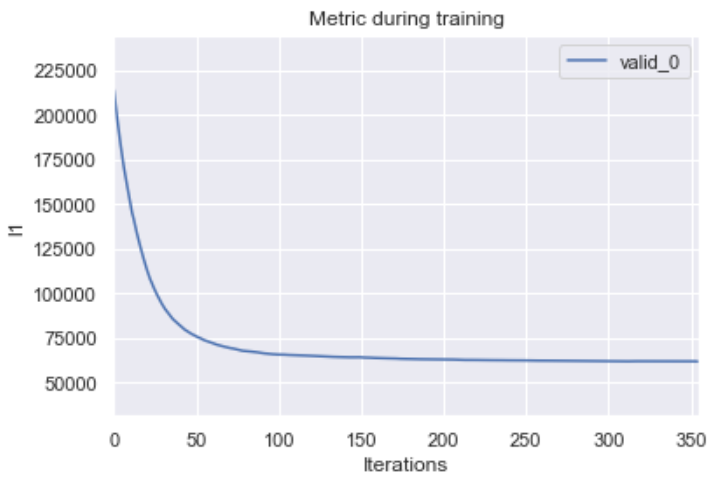


圖9 LightGBM的loss曲線圖

全連階層整合上述三模型 : 透過簡單的模型將DNN、KNN、LightGBM的輸出結果整合，可以看到輸出的loss並無Overfitting。

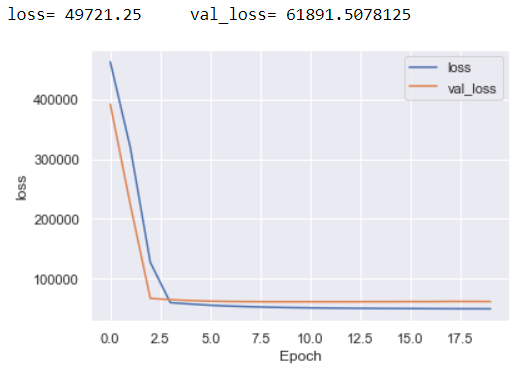


圖10 整合模型的loss曲線圖

1. 結論與改進方向
2. 加深網路，並增加Dropout層及L2正規化抑制Overfitting，嘗試以此方式得到更佳的結果。
3. KNN的做法雖然簡單且直觀，但只看距離就決定結果也過於單調，其與類神經網路雖同為監督式演算法，但KNN無法經過訓練過程學習資料特性，所以在準確度上會有天花板，故未來希望以其他種演算法取代KNN，以不同的方式萃取特徵，以獲得更多更重要的訊息。