

House Prices - Advanced Regression Techniques (group 16)

d09942002 何文劭 b06901146 李世愷 r08246009 許哲維

1. Introduction & Motivation :

這一個資料分析是依據 House Prices: Advanced Regression Techniques 的資料，去預測每個房屋的銷售價格，對於data更詳細的資料請見 Preprocessing/Feature Engineering。

2. Data Preprocessing/Feature Engineering :

在Dataset當中，含有總共 79 項特徵，但其中有許多彼此相關的特徵，因此我們會進行以下的Feature Engineering:

Remove outliers

在練集中特徵對應結果(Sale price)的分佈圖，我們發現會有某些點嚴重偏離出群體之外，因此我們將這些資料點視為outlier進行移除。

LotFrontage	去除 > 300 的資料點
MasVnrArea	去除 > 1,400 的資料點
BsmtFinSF1	去除 > 5,000 的資料點
TotalBsmtSF	去除 > 6,000 的資料點
1stFlrSF	去除 > 4,000 的資料點
GrLivArea	去除 > 4,500 的資料點
BedroomAbvGr	去除 > 8 的資料點
TotRmsAbvGrd	去除 > 14 的資料點
MiscVal	去除 > 8,000 的資料點

Missing data replacing:

根據資料集所附的說明，某些特徵的缺失值(NA)代表的意義可能是該項交易的房屋中並沒有此特徵所代表的設施或物件(如車庫、泳池等)對於這類特徵，我們會將它視為one-hot encoding的其中一項來進行編碼，此外如數值(numerical)的特徵中或在說明中沒有特別表明的特徵缺失值，我們則會進行不同的處理來取代這些缺失值。

A. LotFrontage(Numerical): Linear feet of street connected to property.

由於在相同區域(Neighborhood)會有較高可能有接近大小的街道，因此我們選擇使用相同Neighborhood欄位中的LotFrontage的中位數來取代NA

B. GarageYrBlt(Numerical): Year garage was built

若此房產沒有車庫，則此欄位會被設為NA，為了統一將其設為0

C. MSZoning(Categorical): The general zoning classification

由於相同區域(Neighborhood)可能會較有機會是同一種類，因此我們選擇使用相同Neighborhood欄位中的MSZoning的中位數來取代NA

D. Functional(Categorical): Home functionality

在data_description.txt中表示所有NA應被視為Typ class

E. SaleType(Categorical): Type of sale

使用最多出現的WD class 取代NA

F. MSSubClass(Numerical): Identifies the type of dwelling involved in the sale.

NA可能代表房屋不在其中定義的類別中，因此設置為0

Encoding

在78種特徵中，有些類別(Categorical)特徵屬於類別尺度的調查，因此此類類別的不同類別間有順序關係，我們採用Label encoding來處理以下欄位的資料：

FireplaceQu, BsmtQual, BsmtCond, GarageQual, GarageCond, ExterQual, ExterCond, HeatingQC, PoolQC, KitchenQual, BsmtFinType1, BsmtFinType2, Functional, Fence, BsmtExposure, GarageFinish, LandSlope, LotShape, PavedDrive, CentralAir, MSSubClass, OverallCond, YrSold, MoSold

剩餘的其他的類別特徵資料則採用One-hot encoding來編碼，而數值特徵則採用Data Normalize方式將資料限縮至相近的範圍。

Create New Feature

我們認為年份的資訊對於房價的影響也會是滿大的，因此新增的一些與年份相關的變數，如下：

銷售時屋齡 = $YrSold - YearBuilt$

銷售時屋齡(整修) = $YrSold - YearRemodAdd$

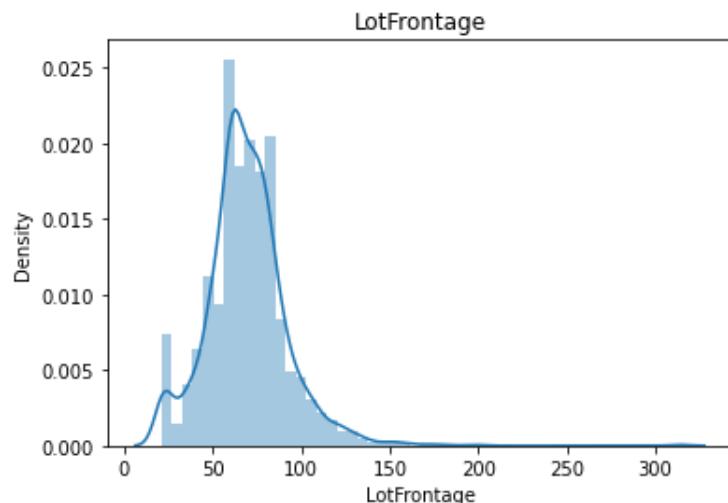
幾年前建造車庫= $YrSold - GarageYrBlt$

Polynomial transform

我們發現特徵中有某幾項與SalePrice有較大的相關度，因此針對這 4 個變數做”平方項”的變化： $MiscVal, 1stFlrSF, 2ndFlrSF, GrLivArea$ ，讓他們在模習中所站的權重比較大。

Skewness

對data用kernel density estimation後我們可以發現有些變數有向左偏的現象，如下圖



在模型當中，當特徵分佈較為接近常態分佈時會有比較好的表現，因此我們針對這類型的變數使用Box-Cox transformation，讓資料向右偏，比較接近常態分佈。

Box-Cox transformation:

$$y(\lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda}, & \text{if } \lambda \neq 0; \\ \log y, & \text{if } \lambda = 0. \end{cases}$$

3. Model Description :

為了避免overfitting 我們有先將 training data 20% 的data抽出作為validation data。

在做完Data Preprocessing後我們會先各別嘗試以下兩大種類的方法

(1) Regression method:

Ridge regression、Lasso、Elastic net、Support vector regression

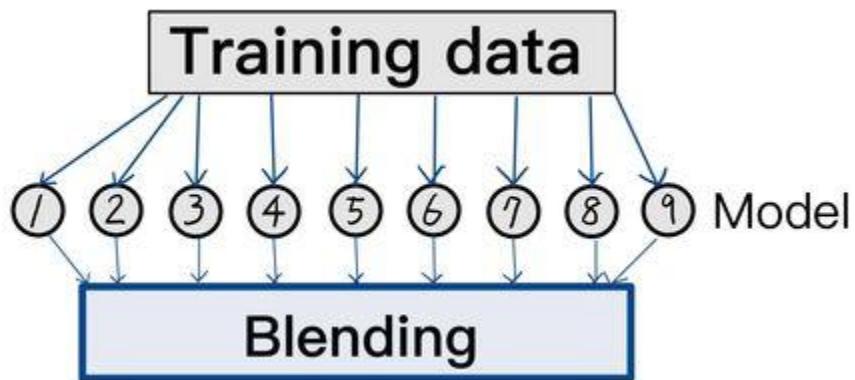
(2) Boosting method

Adaboost、Gradient boosting、XGBoost、Stacking、LightGBM

我們使用XGBoost (為boosting method採用平行運算的版本同時加入L2 regularization term, 以簡化計算複雜性, 再提高精準度) 和 LightGBM (僅選擇使用對增益最大的節點進行深入分解的方法, 如此可以節省了大量分裂節點的資源), 這兩種在眾多比賽中表現優異的方法, 同樣加入我們模型的比較。

因此, 以這9種方法的cross-validation 選擇參數, XGBoos和LightGBMt的參數很多, 所以會使用random search的方式, 選擇cv score最佳的參數, 作為我們的baseline。

最後我們會嘗試ensemble的方式, 將上述得到的6個模型, 直接整合利用blending的方式, 組合出新模型, 與9個baseline比較performance。



4.Experiment and Discussion :

(1) 比較 data preprocessing 對於public score的影響

	Public score
Modify skewness	0.11897
Not modify skewness	0.11980

	Public score
Remove outlier	0.11897
Not remove outlier	0.11980

從上面表格可以看出在model都一樣的架構下，可以發現有做modify skewness 和remove outlier 等data preprocessing的public score會比直接拿原本資料做預測的效果更好。

(2) 比較各種model 的RMSE

Method	RMSE training	RMSE validation
Ridge regression	0.092	0.109
Lasso	0.099	0.107
Elastic net	0.096	0.106
Support vector regression	0.053	0.115
Adaboost	0.094	0.108
Gradient boosting	0.054	0.113
XGBoost	0.084	0.124
Staking	0.097	0.107
LightGBM	0.014	0.131

從上面的表格可以看到boosting類的model training error較小，但validation error 較高(相較於Elastic net、lasso)，可能會有overfitting的問題。

(3) 比較 model 對於public score的影響

	Public score
Full model	0.11897
Linear model	0.13169
Boosting model	0.12718

從表格我們可以看出model blending後表現較好。雖然Boosting model中在validation error較高，但是在public score上還是高於Linear model許多，我們推測是由於Linear regression model對於outlier較為敏感，再加上我們在preprocessing中將部份outlier給移除，因此表現在public score上較差。不過如果我們將兩者進行blending後的結果可以發現能夠比使用單一模型的public score高出許多，推測試blending將不同模型的優缺點進行互補，使model有更進一步的增強。

Kaggle result:

286	xxx_boy		0.11782	39	2mo
287	si_Ro31		0.11782	10	3mo
288	NTU_d09942002_		0.11787	47	15m
Your Best Entry ↑					
Your submission scored 0.12020, which is not an improvement of your best score. Keep trying!					
289	dimasheva1		0.11788	31	1mo

5. Conclusion :

在這次的資料分析中我們了解要讓預測結果更準，data preprocessing是很重要的部分，在各種model的比較中我們也可以看到單用boosting model可能會有overfitting 的問題，但透過validation 以及和其他model做blending可以降低error。

在課堂上的報告做房價預測很常被問到的問題是我們是否會拿這個預測方法買房？我們認為在做這種資料分析時資料收集是重點，必須在想要購買的區域收集資料，另外交通便利生活機能(公園、運動中心)等也是重要的變數必須收集以降低誤差。然後我們認為房價預測也適合應用在賣房，透過我們的方法預測要賣的房子市場價格以方便定價。

6.Reference :

House Prices: Advanced Regression Techniques - Python Ensemble Learning 實做

<https://medium.com/@permoonzz/kaggle-house-prices-advanced-regression-techniques-python-ensemble-learning%E5%AF%A6%E5%81%9A-99f757f4d326>

<https://www.kaggle.com/serigne/stacked-regressions-top-4-on-leaderboard#Data-Processing>