<< ML with Python - Final Project >>

House Pricing by Classification

On-line Report Turn-in Due : **2022/06/19 (Sun.) 11:00 pm**

[期末報告注意事項]：

1. 請將期末報告電子檔以 **ipynb** 副檔名格式上傳至學校的教學平台，檔案名稱如下：

   110-2-Final_Group_OO.ipynb （例如：**110-2-Final_Group_1.ipynb**）

2. 期末報告電子檔內，須註明報告標題以及各組員之科系、年級、學號和姓名。

3. 期末報告缺交和遲交者，不能補交，並以零分計算！

   [建議]： 無論是否能完成所有問題需求，請務必於期限內，上傳期末報告電子檔！

< Problem Description >

Ask a home buyer to describe their dream house, and they probably won't begin with the height of the basement ceiling or the proximity to an east-west railroad. But this final-project's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence.

With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this final project challenges you **to predict a price range indicating how high the house price is**, instead of the final price of each house. Hence, this final project becomes **a classification problem for house pricing** instead of a regression one.

However, the **House-Price dataset** for this final project (referred to **[Ref. 1 : House-Price dataset]**) only provides the exact house-sale prices (i.e., 208500, 181500, … in US dollars). *How to "label" these Sale Prices into different categories for* ***the target vector*** is the very first step for your classification models, and you have to do the ***Exploratory Data Analysis (EDA)*** beforehand (referred to **[Ref. 2]**).

Next, you need to do **Data Pre-processing and Cleansing** (referred to **[Ref. 2, 3 & 4]**) for the features (such as missing values, relabelling, etc.), and apply the **Feature Engineering methods** (referred to **[Ref. 5]**) to find out the appropriate features for the machine learning later.

And then, you should ***train-test-split*** both the feature matrix and target vector, use **the following 8 different classification algorithms for estimator modelling with Scikit-Learn** :

(1) **GaussianNB**
(2) **KNeighborClassifier**
(3) **LogisticRegression**
(4) **DecisionTreeClassifier**
(5) **RandomForestClassifier**
(6) **ExtraTreeClassifier**
(7) **MLPClassifier**
(8) **StackingClassifier**

Moreover, compare the results from the estimators above by their ***accuracy scores*** as well as ***Confusion Matrices*** (referred to **[Ref. 6]**).

If the results (including both accuracy scores and confusion matrices) are not satisfied, then you should go back to "***re-label***" the target, i.e., Sale Prices, again, and repeat all the steps above in order to obtain the desired estimator for classification.

Finally, based on your results and efforts, you should make your conclusion to finish your final-project report.

[Problem] : House Pricing by Classification

[STEP 1] : Download & import the **house_price** dataset.
[STEP 2] : Exploratory Data Analysis & "***label***" the target, **"*SalePrice*"** data.
[STEP 3] : **Data preprocessing & cleansing** for the features.
[STEP 4] : **Feature Engineering** for obtaining the **appropriate features**.
[STEP 5] : Machine learning for building **at least 5 classification models**, and show their **accuracy scores** & **Confusion Matrices**.
If necessary, repeat **STEP 2 ~ 5** *until the results are satisfied.*
[STEP 6] : Discussion.

REFERENCE

1. House-Price dataset - Download : (*data_description.txt* & ***house_price.csv***)
https://drive.google.com/drive/folders/1kCRgFTEqAfrygEbEmQtbW1fb6TZQ-XLu?usp=sharing

[NOTE] : The original dataset is from *Kaggle Competition* - **House Prices - Advanced Regression Techniques** : https://www.kaggle.com/c/house-prices-advanced-regression-techniques

2. Abraham Anderson, **“House Price Competition: Advanced Regression” - Exploratory Data Analysis**, *Kaggle Competition*, 2021/6/16. https://www.kaggle.com/abrahamanderson/house-price-competition-advanced-regression

3. 胡中興老師 教學頻道 , “「**Deep Learning Webinars – Kaggle InClass Competition** 」 **- Kaggle Competition - House Prices - Advanced Regression Techniques: RFs, XGBoost, Stacking, Blending”,**
[YouTube video] : https://youtu.be/fCk4jh0mYWA
[Code] : https://www.kaggle.com/code/lavanyashukla01/how-i-made-top-0-3-on-a-kaggle-competition

4. 胡中興老師 教學頻道 , “「**Python** 資料分析」**- PART II - Pandas Programming”**,
[YouTube video] : https://youtube.com/playlist?list=PLY7ZYxoK-lg6Gw2QkfqF1zEF2CVr5DskP
[Code] : https://drive.google.com/file/d/1bVZrD_AJ89byf2qJUtg8XE--vpJ-rkQM/view

5. 胡中興老師 教學頻道 , “「**Python** 資料分析」**- PART IV - Feature Engineering (特徵工程) ”**,
[YouTube video] : https://youtube.com/playlist?list=PLY7ZYxoK-lg4H1MUg07Lv6JCcFdo898p2
[Code] : https://drive.google.com/file/d/1KsbLuzC14ysRaVnCf27izG03nwzYD1aT/view?usp=sharing

6. Jake VanderPlas, **“Python Data Science Handbook” – Sec. 5.2**, *O’Reilly*, 2017.
https://jakevdp.github.io/PythonDataScienceHandbook/05.02-introducing-scikit-learn.html