大作业：房价预测

目标：利用一系列**可能**与房价相关的因素(特征)训练出一个房价预测模型

---

1. 文件结构:

`data` 文件夹: 存放有：

- `train.csv`: 训练集的特征以及对应的目标预测值(即房价)
- `test.csv`: 测试集的特征
- `test_groundtruth.csv`: 测试集对应的真实房价(即 Y)
- 学号_姓名`.csv`: 预测结果示例，输出预测的结果需要和此文件相同

注意：本次作业给出了测试集的 Y，主要是帮助更好地大家分析预测模型，测试集的数据不应该被用于训练。

本次作业的评分主要依据是提交的代码和实验报告，考察依据是对数据的分析以及利用分析的结果尝试构建一个更好的预测模型，预测模型的效果只作为一个指标上的参考。

数据字典文件夹：存放特征说明文件

`baselines.ipynb`：基线预测模型，可以直接运行，可在此基础上搭建自己的预测模型。

---

2. `xxx.ipynb` 文件如何打开?

`xxx.ipynb` 是 jupyter notebook 文件，可在浏览器中交互式地执行代码，同时可以将代码、文字完美结合起来，在数据科学领域相关（机器学习、数据分析等）被广泛使用。

可以通过以下命令事先安装好 jupyter notebook: `pip install jupyter`

安装好 jupyter 后通过终端进入 xxx.ipynb 所在目录，执行 jupyter notebook 即可在浏览器打开 `xxx.ipynb` 文件

---

3. 预测模型评估指标

所搭建的预测模型将以 MAPE(mean absolute percentage error)作为评估指标，其计算公式可以参考如下链接：

`https://en.wikipedia.org/wiki/Mean_absolute_percentage_error`

可以通过 sklearn 库中的 api 计算，具体操作如下:

```
from sklearn.metrics import mean_absolute_percentage_error
mean_absolute_percentage_error(test_y, pred_y)
```

---

4. 提交与评分

本次作业通过教学网提交，需要包括以下文件：

1. 实验报告：包括但不限于背景介绍，数据探索性分析，特征工程，预测模型的建立，结果的分析与总结等。

注：需在不少于 3 个角度（如特征选择，特征清洗补全，模型算法选择等）对测试结果进行分析和对比，证明所选择方法的合理性和有效性。

2. 源代码：可以是原始 py 文件或者 jupyter notebook 文件，提交的版本需要是干净，易懂，可运行，能输出预测结果的。

3. 预测结果：输出内容需要和`学号_姓名.csv`相同，需要文件命名规则，例如张三，学号 20232023，则预测结果文件命名需为`20232023_张三.csv`

本次作业实验报告占比 80%。预测模型指标占比 20%，其中计算规则如下：

- MAPE 达到 0.3000：20%（指标部分满分）
- MAPE 达到 0.3200：15%
- MAPE 达到 0.3400：10%

本次作业所提供基线模型的 MAPE 为 0.3680。若 MAPE 未达到 0.3400，则预测模型指标的部分得分将按照 (0.3400/your MAPE)*10% 计算。

---

5. 补充材料

【matplotlib】
`https://matplotlib.org/stable/index.html`
【seaborn】
`https://seaborn.pydata.org/`
【sklearn】
`https://scikit-learn.org/`
【pandas】
`https://pandas.pydata.org/docs/`