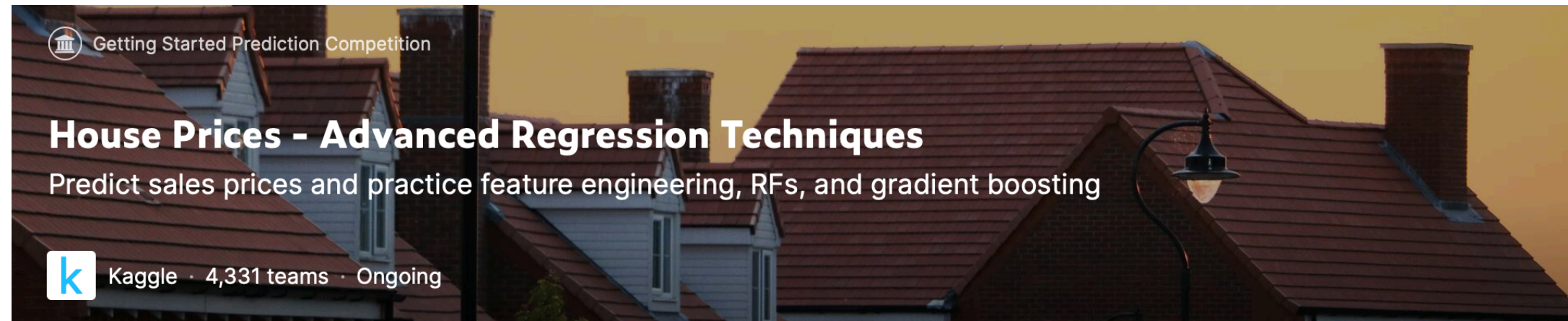


2025大数据导论大作业

比赛介绍



- 利用各方各面的描述变量，预测爱荷华州艾姆斯每套住宅的最终价格。
- 关键技术：特征工程、回归预测
- 比赛网址：<https://www.kaggle.com/competitions/house-prices-advanced-regression-techniques>

数据说明

- 共提供四个文件 train、test、data_description、sample_submission

- train.csv

共1460条记录，每条记录包含住宅的唯一标识ID、79个相关属性及最终成交价格。

| ID | MSSubClass | MSZoning | | SaleCondition | SalePrice |
|------|------------|----------|-------|---------------|-----------|
| 住宅标识 | 建筑类型 | 区域类型 | | 销售条件 | 成交价格 |

- test.csv

共1459条记录，每条记录代表一个待预测住宅的属性。

- data_description.txt

详细介绍每个房屋属性的值域及含义。

- sample_submission.csv

为每个ID预测对应的salePrice。

| ID | SalePrice |
|------|-----------|
| 住宅标识 | 成交价格 |

评价指标

- 使用均方根误差（RMSE）为唯一评价指标

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (\log \hat{y}_i - \log y_i)^2}{N}}$$

* $\log()$ 运算将价位高低不同的预测误差对最终结果的影响统一起来

课程要求

- 每位同学独立完成
- 截止日期：2025年11月12日23时59分59秒
- 提交内容：项目报告、源代码、可执行程序、排行榜排名
- 评分标准：排行榜排名 + 项目报告