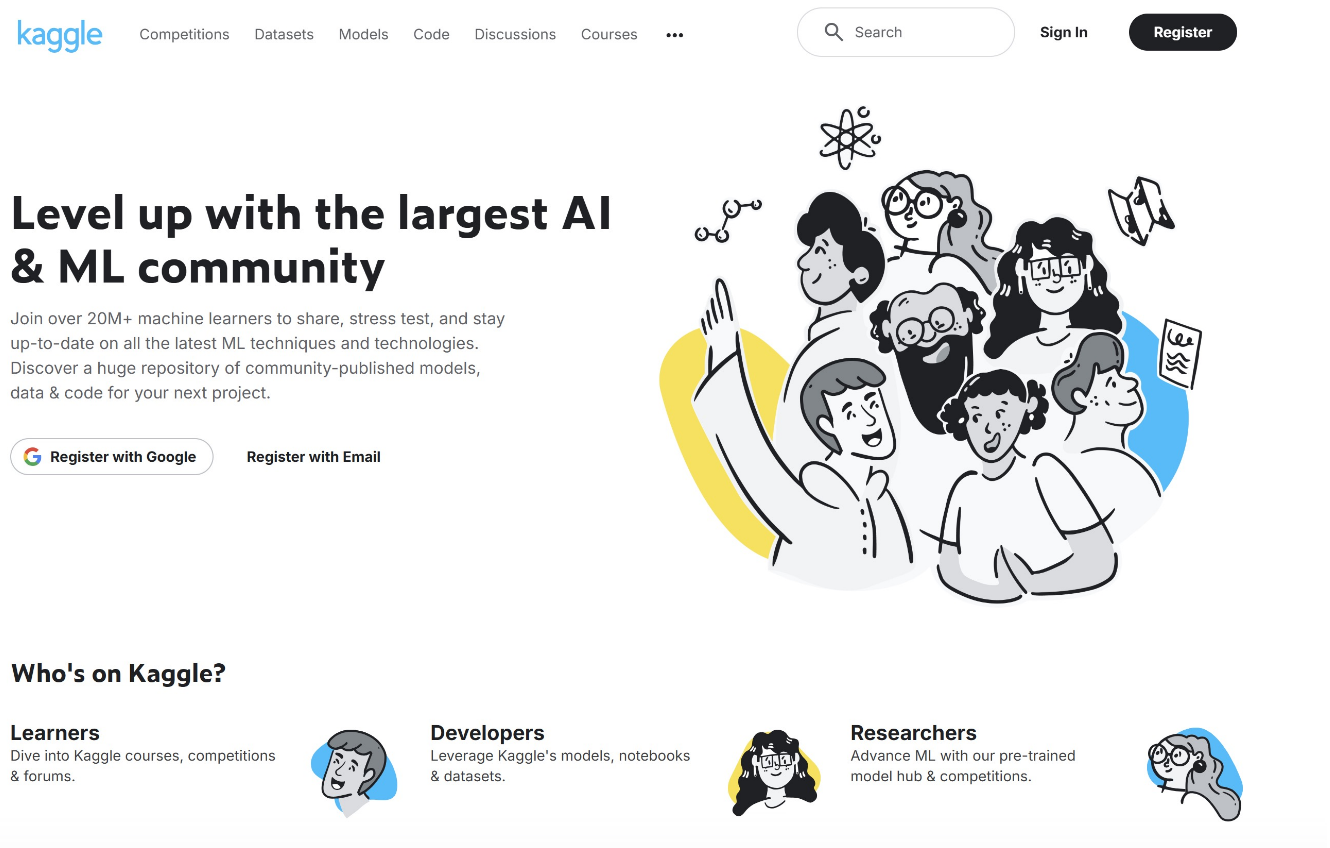
Kaggle实战-房价预测

丁文超

Kaggle

Kaggle:著名的供机器学习爱好者交流的平台

https://www.kaggle.com



Level up with the largest AI & ML community

Join over 20M+ machine learners to share, stress test, and stay up-to-date on all the latest ML techniques and technologies. Discover a huge repository of community-published models, data & code for your next project.

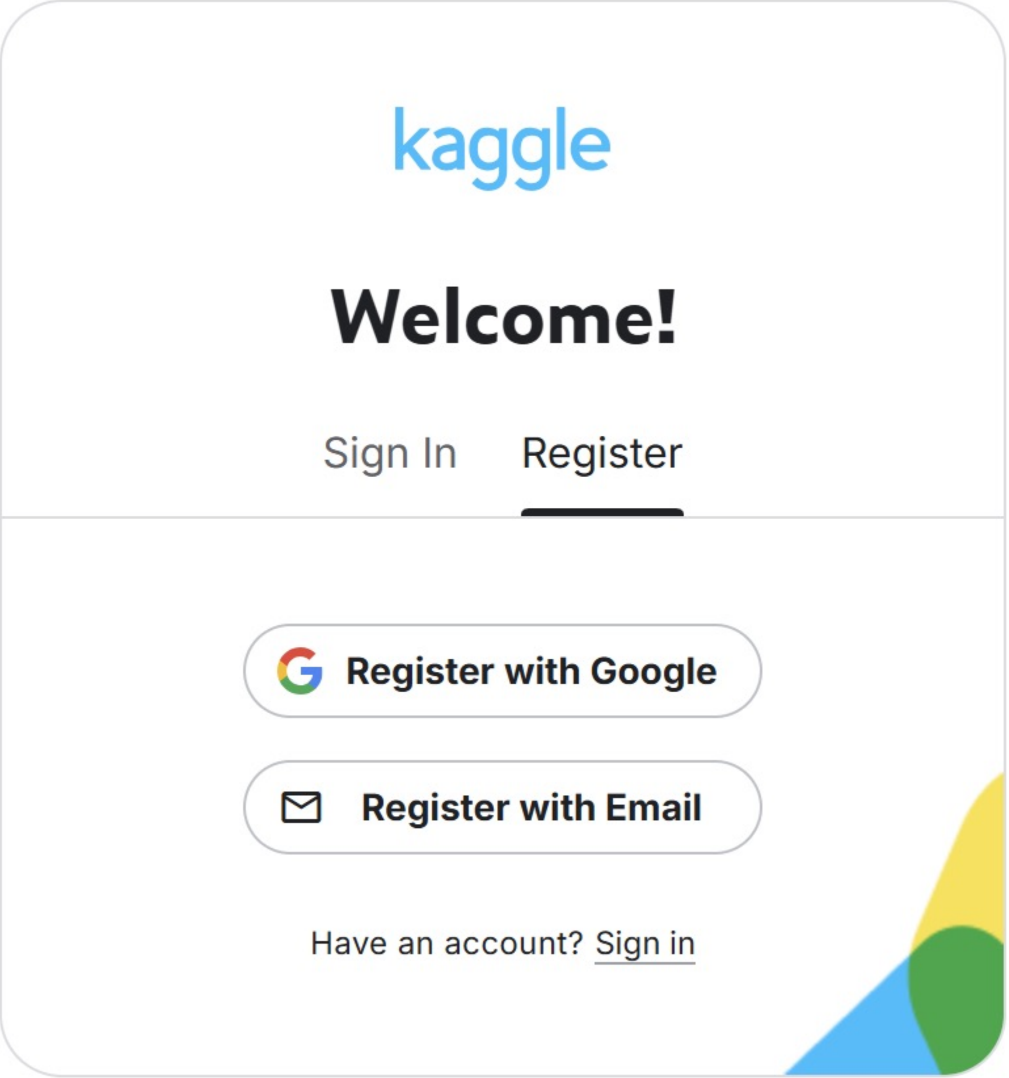
Who's on Kaggle?

Learners
Dive into Kaggle courses, competitions & forums.

Developers
Leverage Kaggle's models, notebooks & datasets.

Researchers
Advance ML with our pre-trained model hub & competitions.

注册账号

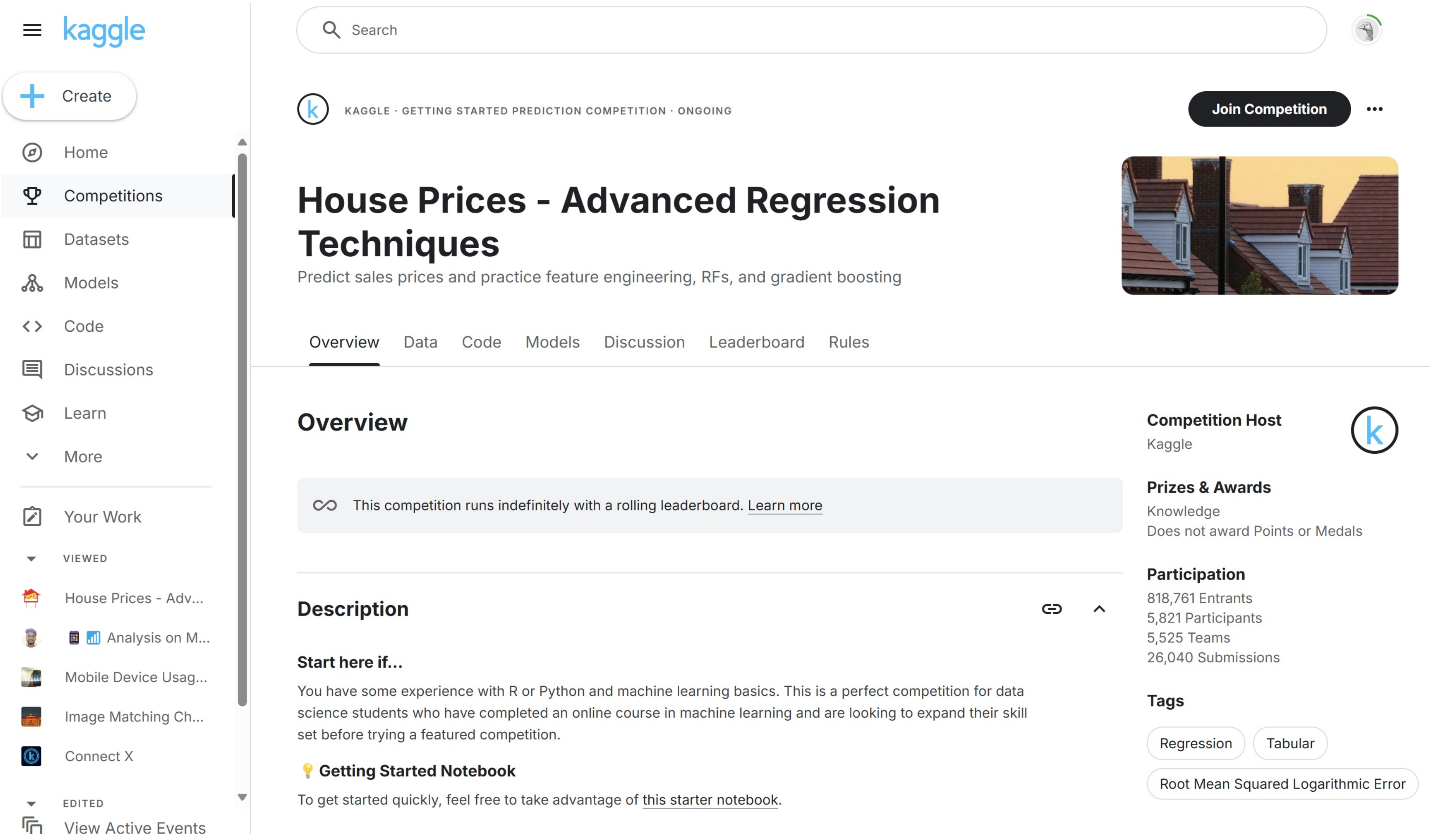


kaggle

Welcome!

无法进行人机验证可以参考 https://blog.csdn.net/sinat_41144773/article/details/103148683

Kaggle实战-房价预测



KAGGLE · GETTING STARTED PREDICTION COMPETITION · ONGOING

House Prices - Advanced Regression Techniques

Predict sales prices and practice feature engineering, RFs, and gradient boosting

Overview

This competition runs indefinitely with a rolling leaderboard. Learn more

Description

Start here if...

You have some experience with R or Python and machine learning basics. This is a perfect competition for data science students who have completed an online course in machine learning and are looking to expand their skill set before trying a featured competition.

💡 **Getting Started Notebook**

To get started quickly, feel free to take advantage of this starter notebook.

Competition Host
Kaggle

Prizes & Awards
Knowledge
Does not award Points or Medals

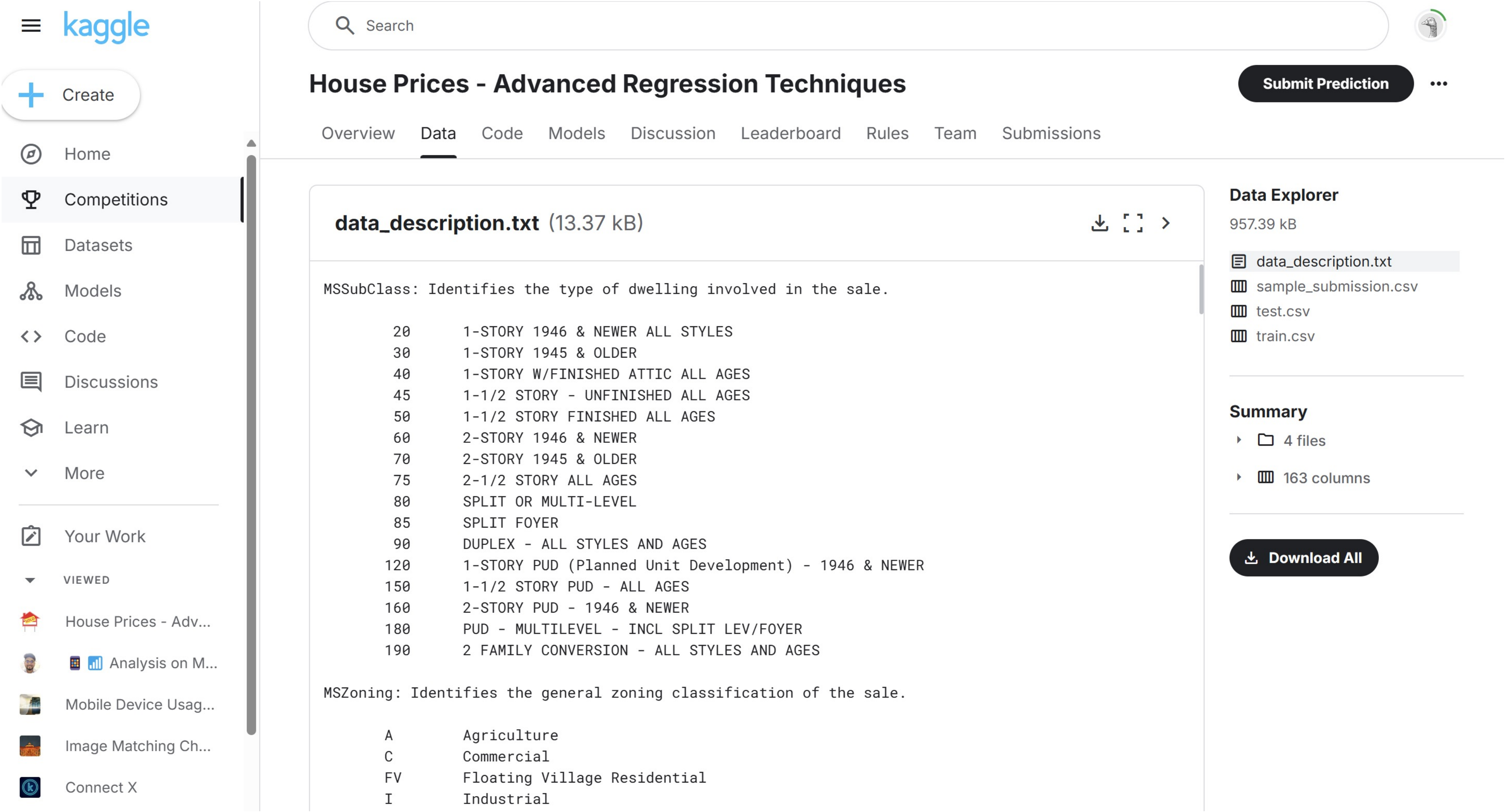
Participation
818,761 Entrants
5,821 Participants
5,525 Teams
26,040 Submissions

Tags
Regression
Tabular
Root Mean Squared Logarithmic Error

网址：https://www.kaggle.com/c/house-prices-advanced-regression-techniques

Kaggle实战-房价预测

下载数据集



```
MSSubClass: Identifies the type of dwelling involved in the sale.

        20      1-STORY 1946 & NEWER ALL STYLES
        30      1-STORY 1945 & OLDER
        40      1-STORY W/FINISHED ATTIC ALL AGES
        45      1-1/2 STORY - UNFINISHED ALL AGES
        50      1-1/2 STORY FINISHED ALL AGES
        60      2-STORY 1946 & NEWER
        70      2-STORY 1945 & OLDER
        75      2-1/2 STORY ALL AGES
        80      SPLIT OR MULTI-LEVEL
        85      SPLIT FOYER
        90      DUPLEX - ALL STYLES AND AGES
       120      1-STORY PUD (Planned Unit Development) - 1946 & NEWER
       150      1-1/2 STORY PUD - ALL AGES
       160      2-STORY PUD - 1946 & NEWER
       180      PUD - MULTILEVEL - INCL SPLIT LEV/FOYER
       190      2 FAMILY CONVERSION - ALL STYLES AND AGES

MSZoning: Identifies the general zoning classification of the sale.

       A        Agriculture
       C        Commercial
       FV       Floating Village Residential
       I        Industrial
```

Kaggle实战-房价预测

比赛数据集

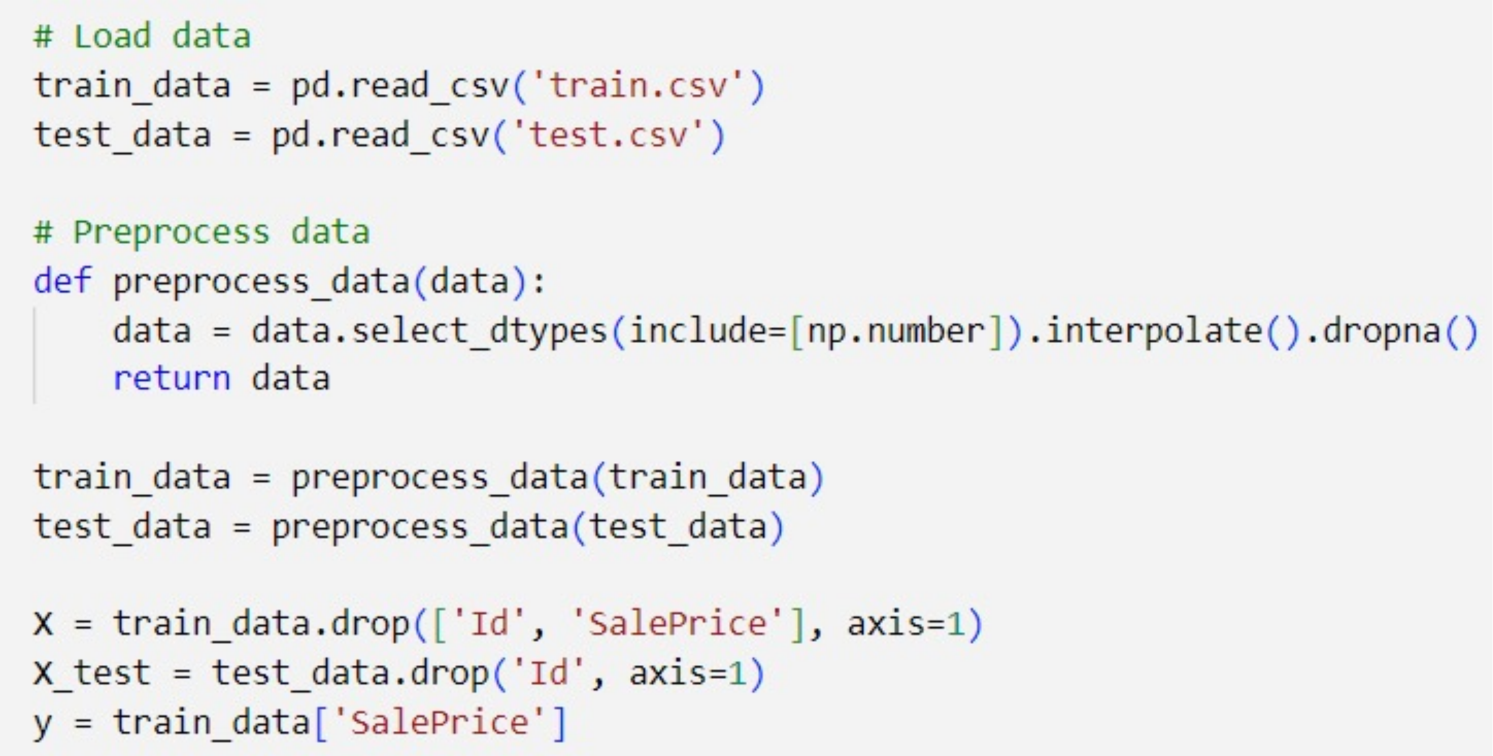
- ➢ 分为训练数据集和测试数据集
- ➢ 训练、测试数据集都包括每栋房子的特征
  - 街道类型
  - 建造年份
  - 房顶类型
  - 地下室状况等
- ➢ 特征值有连续的数字、离散的标签甚至是缺失值“na”
- ➢ 只有训练数据集包括了每栋房子的价格即标签

Kaggle实战-房价预测-样例

读取数据集

- ◆ 数据集第一个特征是id，帮助模型记住每个训练样本，但难以推广到测试样本，所以不使用它来训练

- ◆ 将其他形式为数字的特征提取出来作为输入

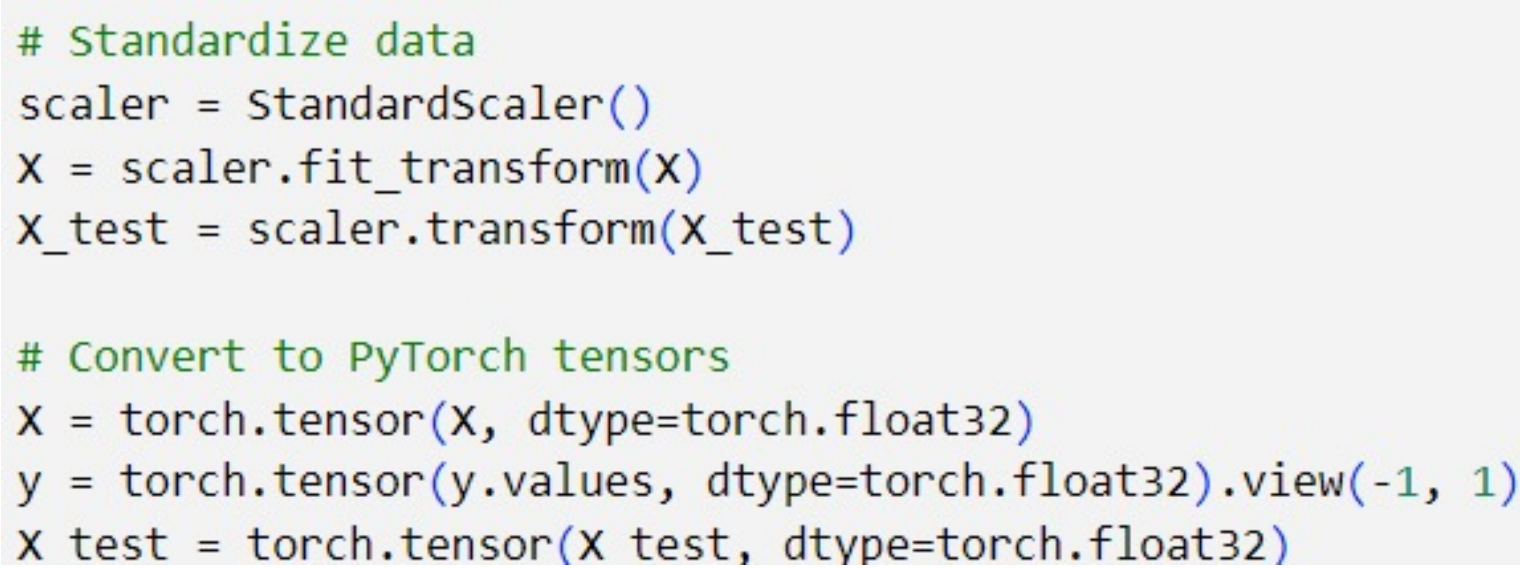


```python
# Load data
train_data = pd.read_csv('train.csv')
test_data = pd.read_csv('test.csv')

# Preprocess data
def preprocess_data(data):
    data = data.select_dtypes(include=[np.number]).interpolate().dropna()
    return data

train_data = preprocess_data(train_data)
test_data = preprocess_data(test_data)

X = train_data.drop(['Id', 'SalePrice'], axis=1)
X_test = test_data.drop('Id', axis=1)
y = train_data['SalePrice']
```

Kaggle实战-房价预测-样例

预处理数据

- ◆ 对特征做标准化
- ◆ 通过values属性转成torch. tensor格式的数据

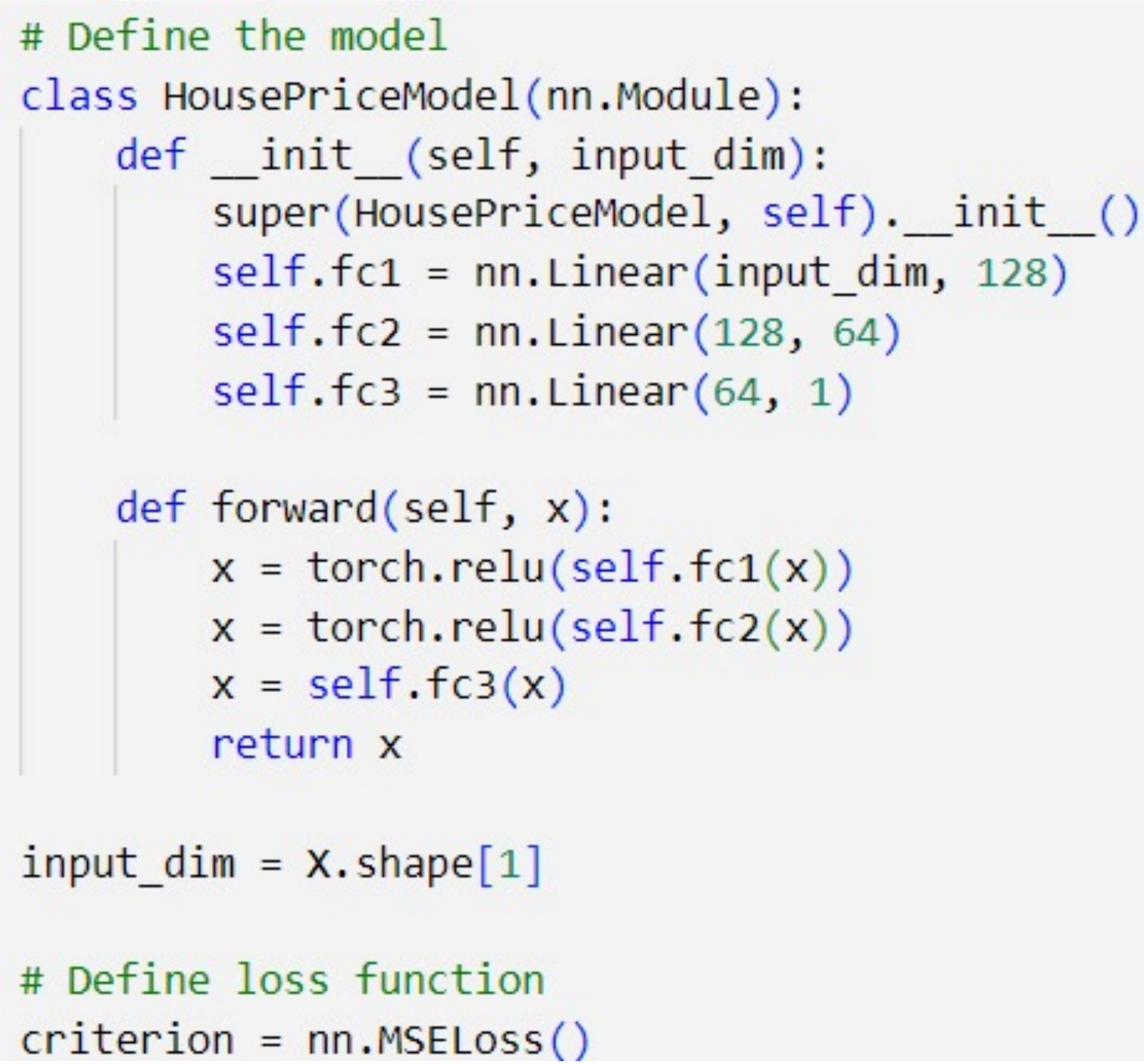


```
# Standardize data
scaler = StandardScaler()
X = scaler.fit_transform(X)
X_test = scaler.transform(X_test)

# Convert to PyTorch tensors
X = torch.tensor(X, dtype=torch.float32)
y = torch.tensor(y.values, dtype=torch.float32).view(-1, 1)
X_test = torch.tensor(X_test, dtype=torch.float32)
```

Kaggle实战-房价预测-样例

训练模型

- ◆ MLP网络
- ◆ 使用对数均方根评价模型



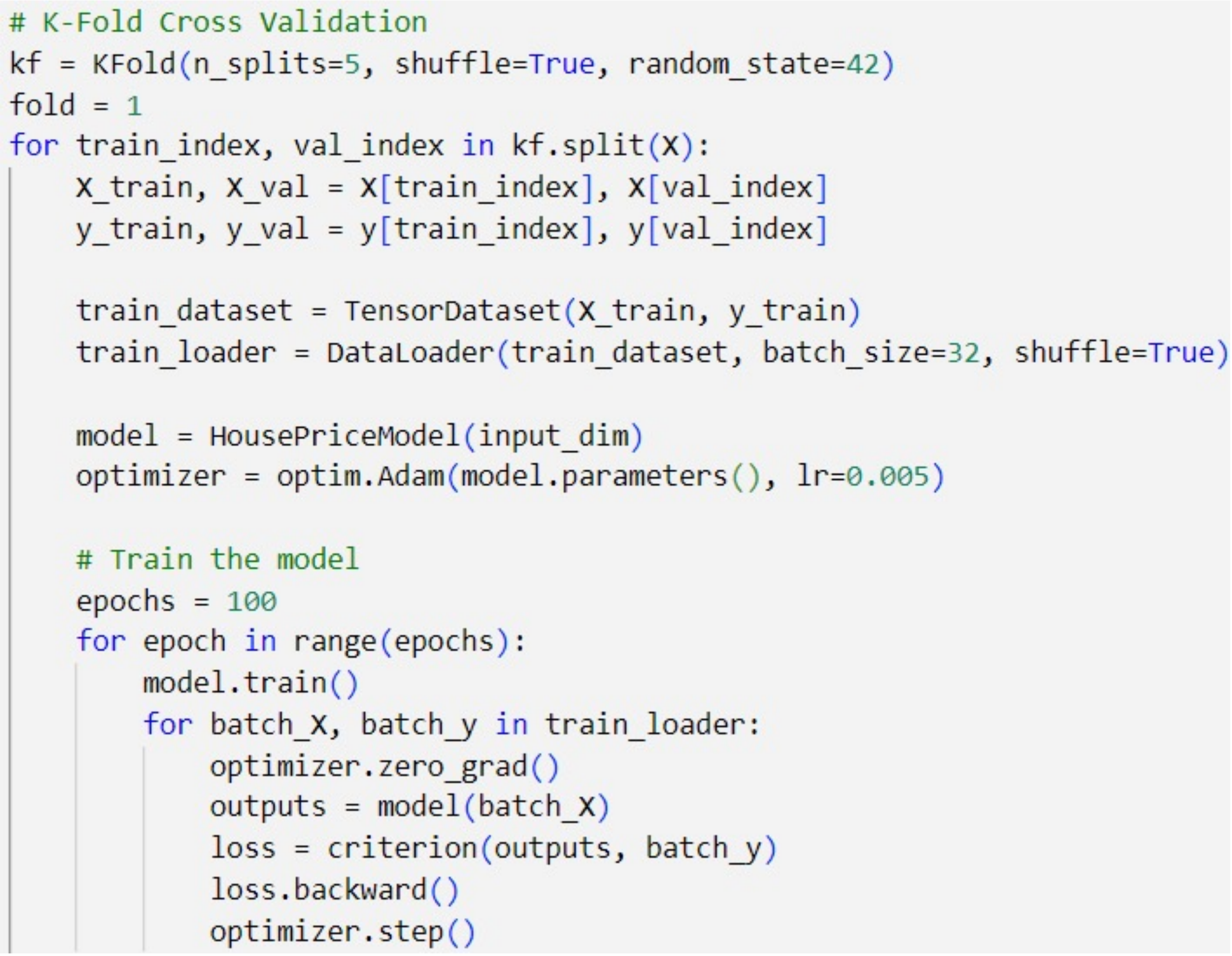
```
# Define the model
class HousePriceModel(nn.Module):
    def __init__(self, input_dim):
        super(HousePriceModel, self).__init__()
        self.fc1 = nn.Linear(input_dim, 128)
        self.fc2 = nn.Linear(128, 64)
        self.fc3 = nn.Linear(64, 1)

    def forward(self, x):
        x = torch.relu(self.fc1(x))
        x = torch.relu(self.fc2(x))
        x = self.fc3(x)
        return x

input_dim = X.shape[1]

# Define loss function
criterion = nn.MSELoss()
```

Kaggle实战-房价预测-样例

K折交叉验证

◆ K折交叉验证用来选择模型设计并调节超参数

◆ 训练K次并返回训练和验证的平均误差



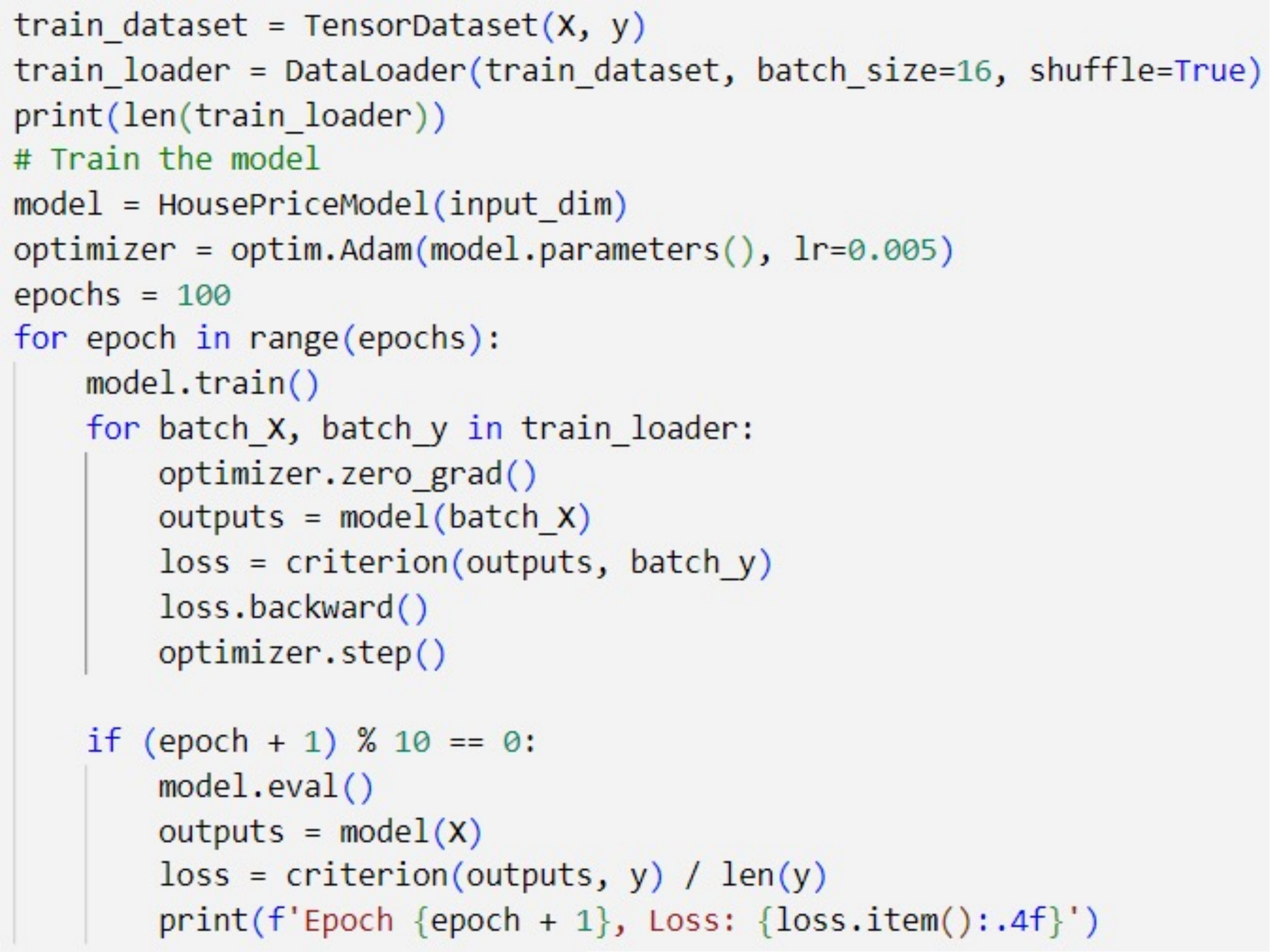
```
# K-Fold Cross Validation
kf = KFold(n_splits=5, shuffle=True, random_state=42)
fold = 1
for train_index, val_index in kf.split(X):
    X_train, X_val = X[train_index], X[val_index]
    y_train, y_val = y[train_index], y[val_index]

    train_dataset = TensorDataset(X_train, y_train)
    train_loader = DataLoader(train_dataset, batch_size=32, shuffle=True)

    model = HousePriceModel(input_dim)
    optimizer = optim.Adam(model.parameters(), lr=0.005)

    # Train the model
    epochs = 100
    for epoch in range(epochs):
        model.train()
        for batch_X, batch_y in train_loader:
            optimizer.zero_grad()
            outputs = model(batch_X)
            loss = criterion(outputs, batch_y)
            loss.backward()
            optimizer.step()
```

Kaggle实战-房价预测-样例

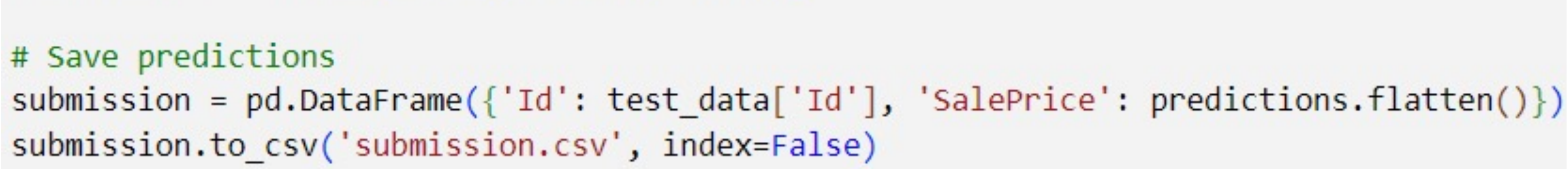
预测

- ◆ 使用完整的训练数据集来重新训练模型

- ◆ 将预测结果存成提交所需要的格式



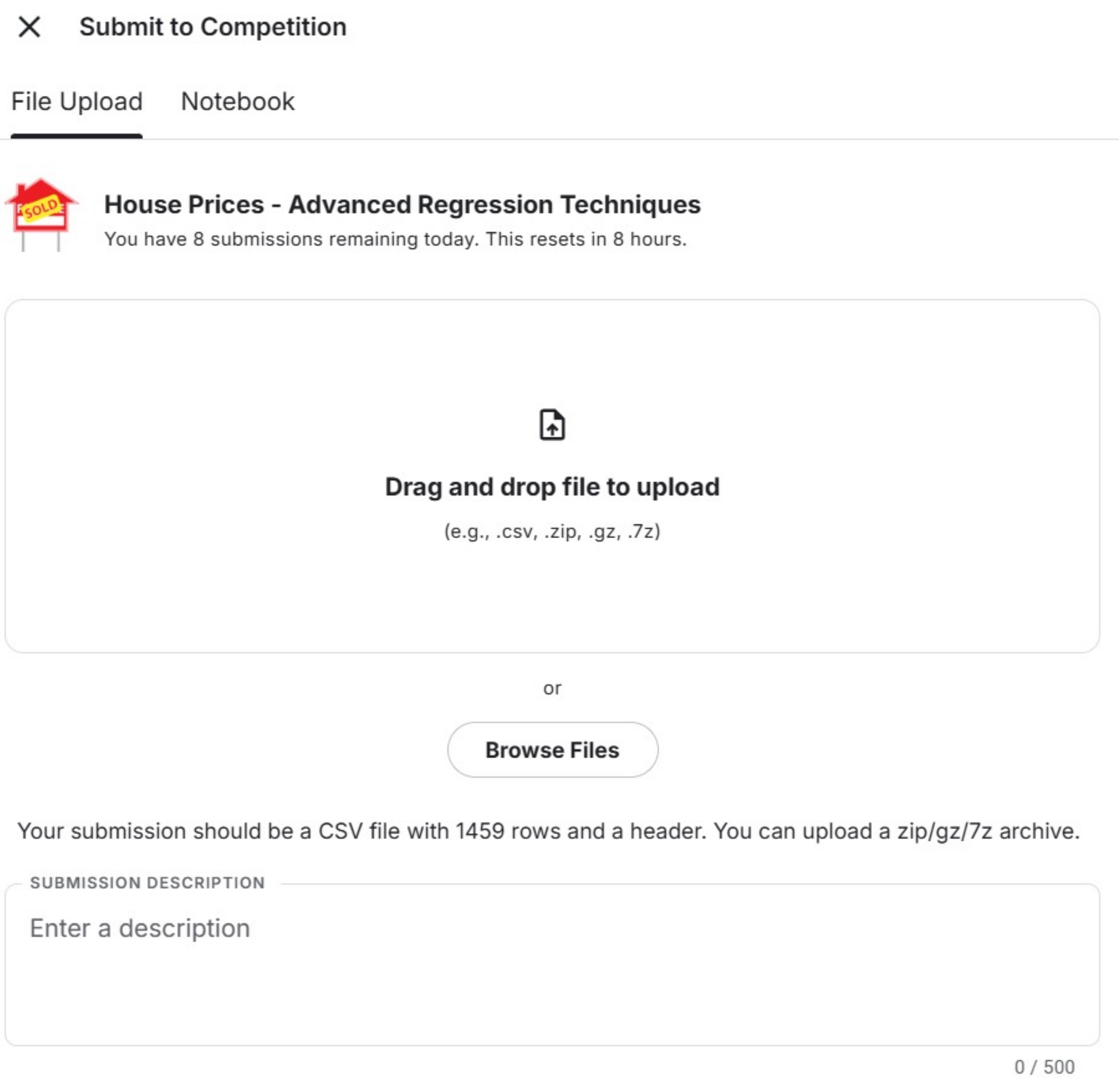
```python
train_dataset = TensorDataset(X, y)
train_loader = DataLoader(train_dataset, batch_size=16, shuffle=True)
print(len(train_loader))
# Train the model
model = HousePriceModel(input_dim)
optimizer = optim.Adam(model.parameters(), lr=0.005)
epochs = 100
for epoch in range(epochs):
    model.train()
    for batch_X, batch_y in train_loader:
        optimizer.zero_grad()
        outputs = model(batch_X)
        loss = criterion(outputs, batch_y)
        loss.backward()
        optimizer.step()

    if (epoch + 1) % 10 == 0:
        model.eval()
        outputs = model(X)
        loss = criterion(outputs, y) / len(y)
        print(f'Epoch {epoch + 1}, Loss: {loss.item():.4f}')
```



```python
# Make predictions on the test set
model.eval()
predictions = model(X_test).detach().numpy()

# Save predictions
submission = pd.DataFrame({'Id': test_data['Id'], 'SalePrice': predictions.flatten()})
submission.to_csv('submission.csv', index=False)
```

Kaggle实战-房价预测

提交预测



Submit to Competition

File Upload   Notebook

House Prices - Advanced Regression Techniques

You have 8 submissions remaining today. This resets in 8 hours.

Drag and drop file to upload

(e.g., .csv, .zip, .gz, .7z)

or

Browse Files

Your submission should be a CSV file with 1459 rows and a header. You can upload a zip/gz/7z archive.

SUBMISSION DESCRIPTION

Enter a description

0 / 500

Kaggle实战-房价预测

你可以采取的方法：

1. 数据处理部分将连续数值、离散数值（文本）分别处理
2. 采用不同的模型架构
3. 通过K折交叉验证来调整模型和超参

THANKS