

哈尔滨工业大学（深圳）

**Harbin Institute of Technology
(Shenzhen)**

项目名称: 大数据导论大作业 2

姓名: 刘睿

学号: 220110720

日期: 2024/11/12

1.项目内容

在本次项目中，我们使用 Kaggle 上的经典数据集 "House Prices: Advanced Regression Techniques" 进行房价预测。目标是建立一个回归模型，预测房屋的最终销售价格。

本项目的数据集包括多个与房价相关的特征，例如房屋的面积、房间数量、地段、建造年份等。数据集分为训练集和测试集，训练集用于模型训练和验证，测试集用于评估模型的泛化能力。

2.项目实施

A. 数据预处理

在数据预处理中，首先对数据进行了缺失值处理和数据编码：

缺失值处理：使用 `SimpleImputer` 对数值型和分类特征进行均值填充和众数填充。

编码：对类别特征使用了 `One-Hot` 编码，以便模型能够处理非数值特征。

标准化：使用 `StandardScaler` 对数值特征进行标准化，以确保特征之间具有相似的尺度。

B. 模型选择与训练

我们使用了多种回归模型，包括 `Gradient Boosting Regressor` 和 `XGBoost Regressor`。通过使用 `GridSearchCV`，对模型的超参数进行了调优，以便找到最优参数组合。

`Gradient Boosting Regressor`：使用了梯度提升回归模型，并调整了学习率、树的深度等参数，得到了一个性能较好的基准模型。

`XGBoost Regressor`：使用了 `XGBoost` 来进一步提升模型的性能，并通过网格搜索确定了最优参数

3.项目结果和分析

模型评估指标主要是均方根误差（RMSE）。通过对验证集的评估，我们发现 `XGBoost` 模型在预测精度上优于梯度提升模型。

在验证集上，我们的 `XGBoost` 模型的 RMSE 达到了一个较低的数值，表明模型具有较好的预测能力。


```
In [17]: # 预测与评估
y_pred = model.predict(X_val)
mse = mean_squared_error(y_val, y_pred)
rmse = np.sqrt(mse)
print(f"验证集 RMSE: {rmse}")
```

验证集 RMSE: 25349.228174458935

在 Kaggle 上最终提交的成绩如下：

2053


Poi Hakurei



0.13906

7

39m



Your Best Entry!
Your most recent submission scored 0.13906, which is the same as your previous score. Keep trying!

4.总结

本项目通过对房价数据的深入分析和特征工程，建立了有效的房价预测模型。我们比较了多种模型的表现，最终选择了 XGBoost 模型作为最优方案。未来，我们可以通过引入更多高级特征工程手段和更复杂的模型结构来进一步提升预测精度。