

课程设计报告

（ 2021秋季学期 ）

|  |  |
| --- | --- |
| 课程名称 | 机器学习课程设计 |
| **项目名称** | **House Prices**  **- Advanced Regression Techniques** |

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名 | 熊方莹 | 学号 | 20190440803 |

|  |  |  |  |
| --- | --- | --- | --- |
| 专业 | 软件工程 | 班级 | 1903班 |

|  |  |  |  |
| --- | --- | --- | --- |
| 地点 | 崇业楼206 | 教师 | 刘永彬 |

1. **项目概述及计划**
2. **问题描述**

使用python或R语言，将项目提供的79个解释变量进行分析，预测得到爱荷华州埃姆斯住宅的每套房屋的最终房价。

1. **模型建模**

首先将数据进行分析，将数据进行清洗后，导入分析模型。将lightGBM，XGBoost，StackedRegressor三种方法的模型数据加权平均后得到最终结果。

1. **实验分析**

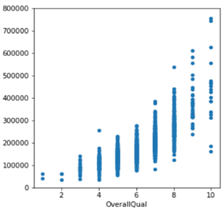
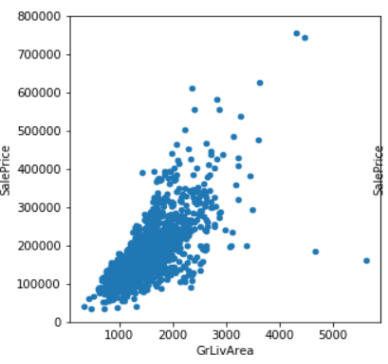
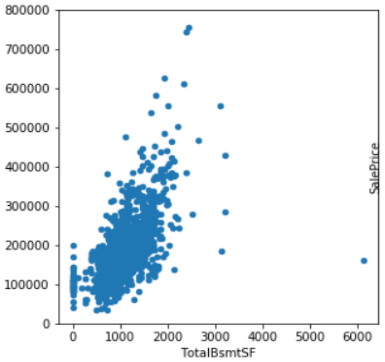
在训练数据集中，存在许多空白数据，所以首先需要对数据进行清洗。实验数据包含79个解释变量，为了是的数据能够更好的拟合结果，需要对数据集进行特征工程属性的分析。得到特征工程后，需要将特征工程与实际数据进行对比，将特征工程属性，按照权重进行比对，最后建模分析。我结合网上的两篇博客对数据进行分析建模。

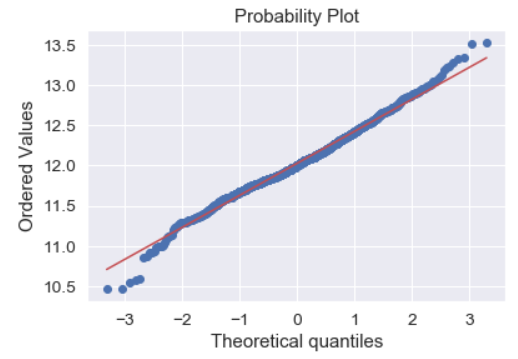
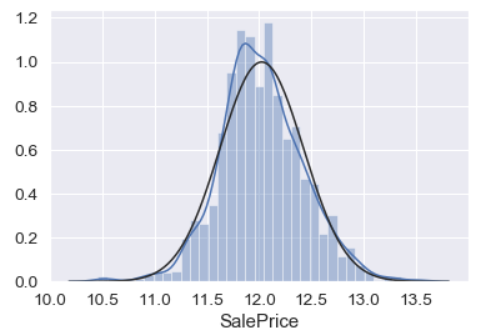
（博客1：https://blog.csdn.net/u012063773/article/details/79349256）

（博客2：https://blog.csdn.net/chandelierds/article/details/83627060）

**4.1 数据分析**

对数据进行图形化展示，然后进行数据清洗。

****

****

**var3 = 'YearBuilt'**

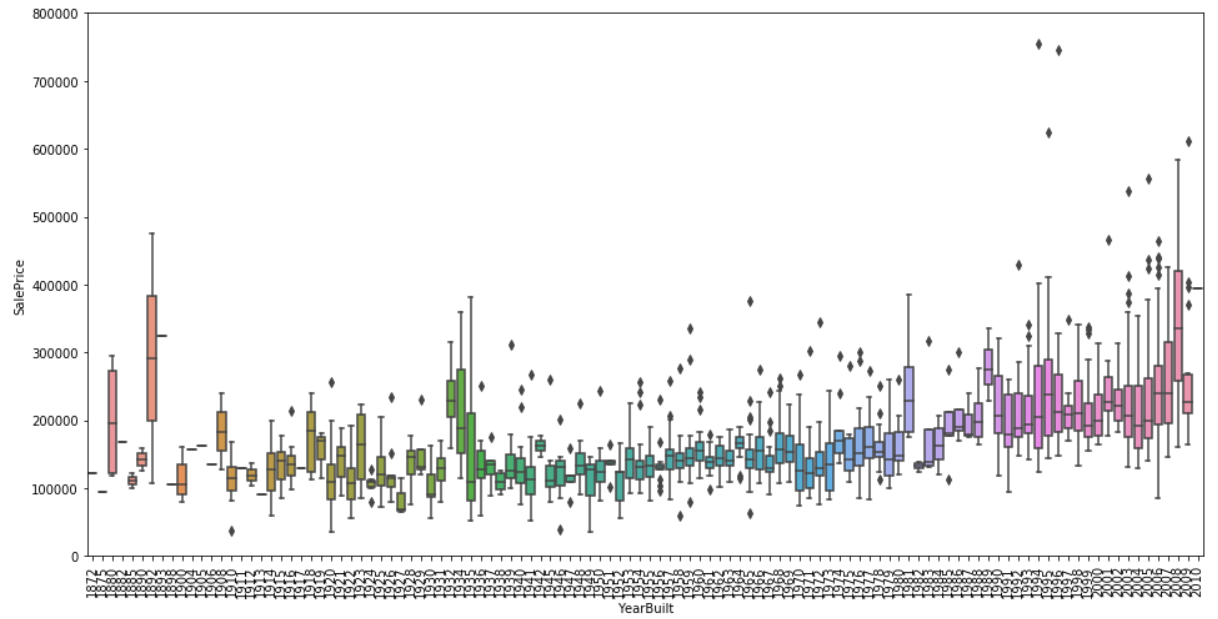
**fig, ax = plt.subplots(figsize=(16,8))**

**sns.boxplot(x=var3,y=output,data=train)**

**ax.set\_ylim(0,800000)**

**plt.xticks(rotation=90)**

**plt.show()**

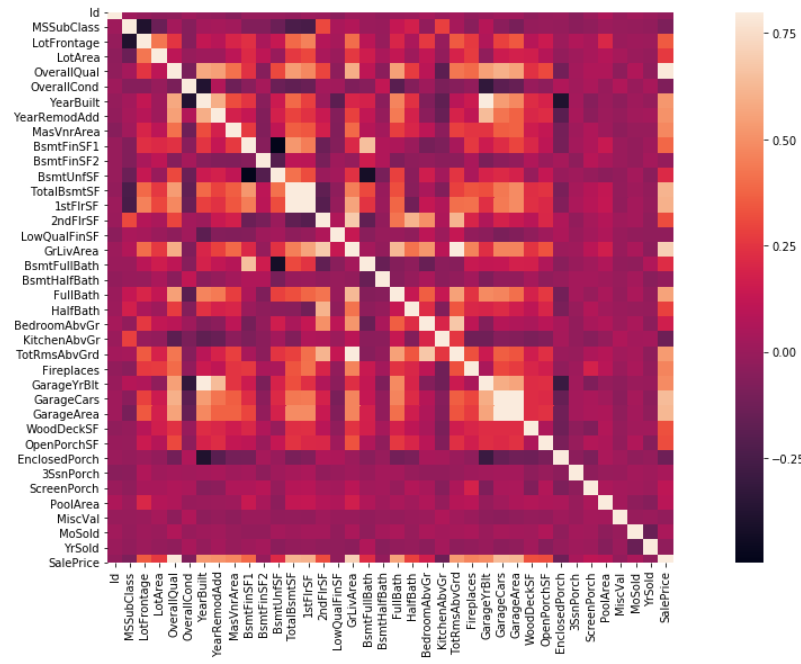
****

**corrmat = train.corr()**

**f, ax = plt.subplots(figsize=(20, 9))**

**sns.heatmap(corrmat, vmax=0.8, square=True, ax=ax) # square参数保证corrmat为非方阵时，图形整体输出仍为正方形**

**plt.show()**

****

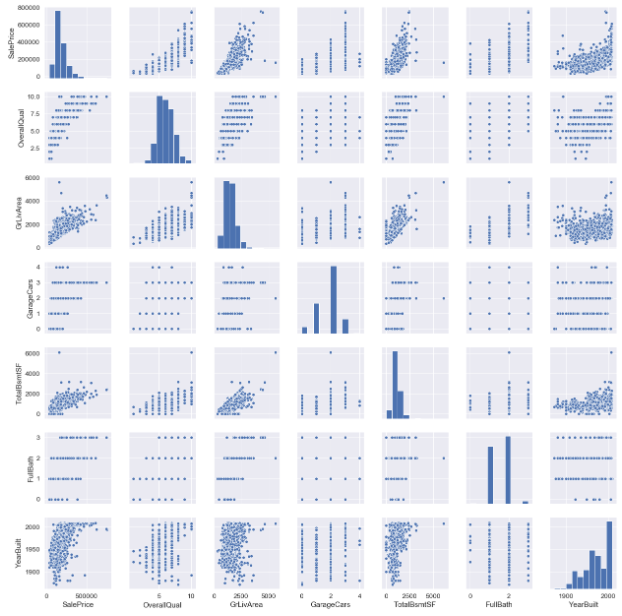
**var\_set=['SalePrice','OverallQual','GrLivArea','GarageCars','TotalBsmtSF','FullBath','YearBuilt']**

**sns.set(font\_scale=1.25) # 设置横纵坐标轴的字体大小**

**sns.pairplot(train[var\_set]) # 7\*7图矩阵**

**# 可在kind和diag\_kind参数下设置不同的显示类型，此处分别为散点图和直方图，还可以设置每个图内的不同类型的显示**

**plt.show()**

****

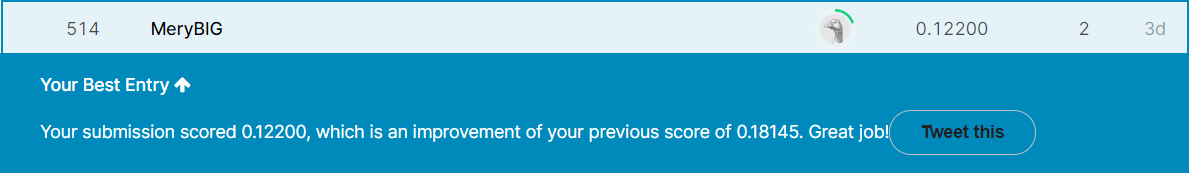
**4.2 实验环境**

**装有Ananconda3以及jupyter notebook的电脑**

**4.3 实验结果**

****

**4.4 实验对比（竞赛排名）**



1. **项目总结**

数据分析是预测的基础部分，否则将对预测造成严重的影响，需深刻分析并挖掘数据之间的关系。模型的选择和调参是机器学习的关键部分，是一大难点，也是一次挑战和学习的机会。

掌握了预测的基本流程，多种模型配合比单一模型效果更加好。积极实操，查阅资料，不断试错，才能有所进步。

模型方面还可以将多个模型组合起来，进行模型融合。