北京大学
PEKING UNIVERSITY

统计分析与商务智能课程项目报告

基于多元回归的房价预测模型

组员：陈麒先 2001213040

组员：赵浩森 2001213039

组员：熊家欣 2001213047

学院：信息科学技术学院

2020 年 12 月

郑重声明

本项目由本小组成员合作完成。抄袭行为在任何情况下都是不能容忍的（COPY is strictly prohibited under any circumstances）！由抄袭所产生的一切后果由抄袭者承担，勿谓言之不预也。

陈麒先

一、项目背景

让一个购房者选购他们心目中理想的房子，他们可能不会喜欢那些位于地下室或者顶层的住房，也不会喜欢靠近铁路或是城市边缘的房子。但观测数据证明，多种变量可能对房屋的最终成交价格产生巨大影响。本项目将根据给定的训练数据集中提供的 79 个解释变量所描述的爱荷华州艾姆斯市住宅的多个特征，建立多元回归模型，预测测试集中给出的每个住宅的最终价格。预测结果将根据预测值的对数和实际销售价格的对数之间的均方根误差(RMSE)进行评估。使用对数意味着，在预测昂贵房屋和廉价房屋时出现的误差，对结果的影响是一样的。

本项目主要用到的数据分析工具为 SAS，SAS（全称 STATISTICAL ANALYSIS SYSTEM，简称 SAS）是全球最大的私营软件公司之一，是由美国北卡罗来纳州立大学 1966 年开发的统计分析软件。1976 年 SAS 软件研究所（SAS INSTITUTE INC）成立，开始进行 SAS 系统的维护、开发、销售和培训工作。期间经历了许多版本，并经过多年来的完善和发展，SAS 系统在国际上已被誉为统计分析的标准软件，在各个领域得到广泛应用。因此，SAS 的引入和应用，将为本项目的具体实施带来巨大便利。

二、数据来源

本项目采用由 Dean De Cock 收集整理的艾姆斯住房数据集，该数据集使用 79 个特征描述了爱荷华州艾姆斯市住宅信息，以及房屋的具体售价。

在 Kaggle 竞赛官网[1]可以下载完整的数据集，其中包括以下四个文件：

- train.csv – 训练集
- test.csv – 测试集
- sample_submission.csv – 样例提交结果
- data_description.txt – 数据描述文件

三、数据描述

本项目的数据集包含以下变量：

- SalePrice：房屋售价，最终预测目标
- MSSubClass：建造等级
- MSZoning：整体区域分类
- LotFrontage：房屋到相连街道的距离
- LotArea：阳台面积
- Street：连接道路类型
- Alley：连接通道类型
- LotShape：房屋整体形状
- LandContour：房屋平整度
- Utilities：可用设施数
- LotConfig：阳台结构
- LandSlope：房屋的草坪
- Neighborhood：在城市中所处位置
- Condition1：到铁路或主公路的临近性
- Condition2：到铁路或主公路的临近性（如果存在第二条路）
- BldgType：居住类型
- HouseStyle：住房风格
- OverallQual：材料和质量
- OverallCond：环境整体评级
- YearBuilt：建造日期
- YearRemodAdd：翻修日期

- RoofStyle: 房顶类型
- RoofMatl: 房顶材料
- Exterior1st: 房屋外部装饰
- Exterior2nd: 房屋外部装饰（如果还有其他外部装饰）
- ExterQual: 外部材料质量
- ExterCond: 外部材料呈现效果
- Foundation: 地板类型
- BsmtQual: 地下室高度
- BsmtCond: 地下室整体情况
- BsmtExposure: 地下室墙体高度
- BsmtUnfSF: 未完成建造的地下室面积
- TotalBsmtSF: 地下室总面积
- Heating: 供暖方式
- HeatingQC: 供暖质量和环境
- CentralAir: 中央空调
- Electrical: 电力系统情况
- 1stFlrSF: 一层面积
- 2ndFlrSF: 二层面积
- LowQualFinSF: 建造质量低的总面积
- GrLivArea: 地上部分使用面积
- Bedroom: 地上部分卧室个数
- Kitchen: 厨房个数
- KitchenQual: 厨房质量

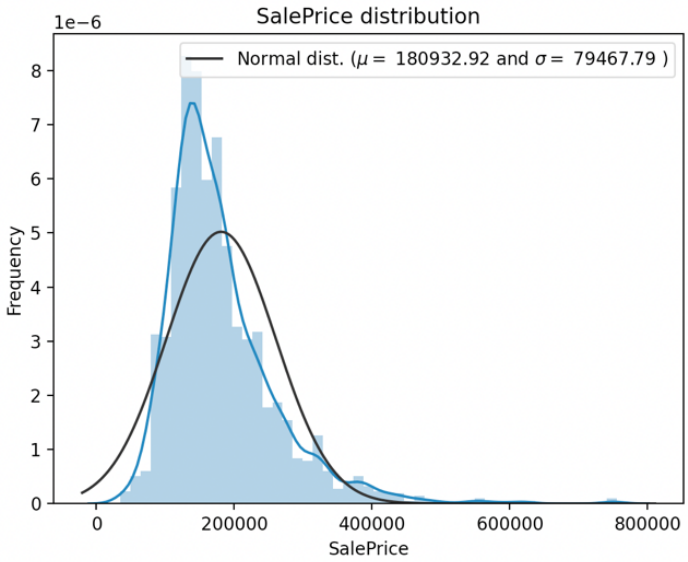
- TotRmsAbvGrd: 不含卫生间的地上部分房间总数
- Functional: 房屋功能评分
- Fireplaces: 火炉数量
- FireplaceQu: 火炉质量
- GarageType: 车库位置
- GarageYrBlt: 车库建造年份
- GarageFinish: 车库建造质量
- GarageCars: 车库容量
- GarageArea: 车库面积
- PoolArea: 泳池面积
- PoolQC: 泳池质量
- Fence: 围栏质量
- MoSold: 销售月份
- YrSold: 销售年份
- SaleType: 销售类型
- SaleCondition: 销售状况

数据集中的每一条数据记录表示的是每间房屋的相关信息，其中训练数据和测试数据分别各有 1460 条，数据的特征列有 79 个，其中 35 个是数值类型的，44 个类别类型。

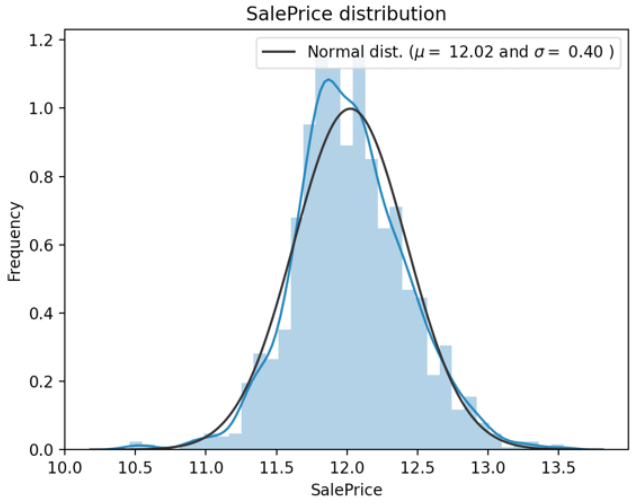
四、数据处理

1、数据探索

下载好数据集后，打开训练数据集 train.csv 进行观察。

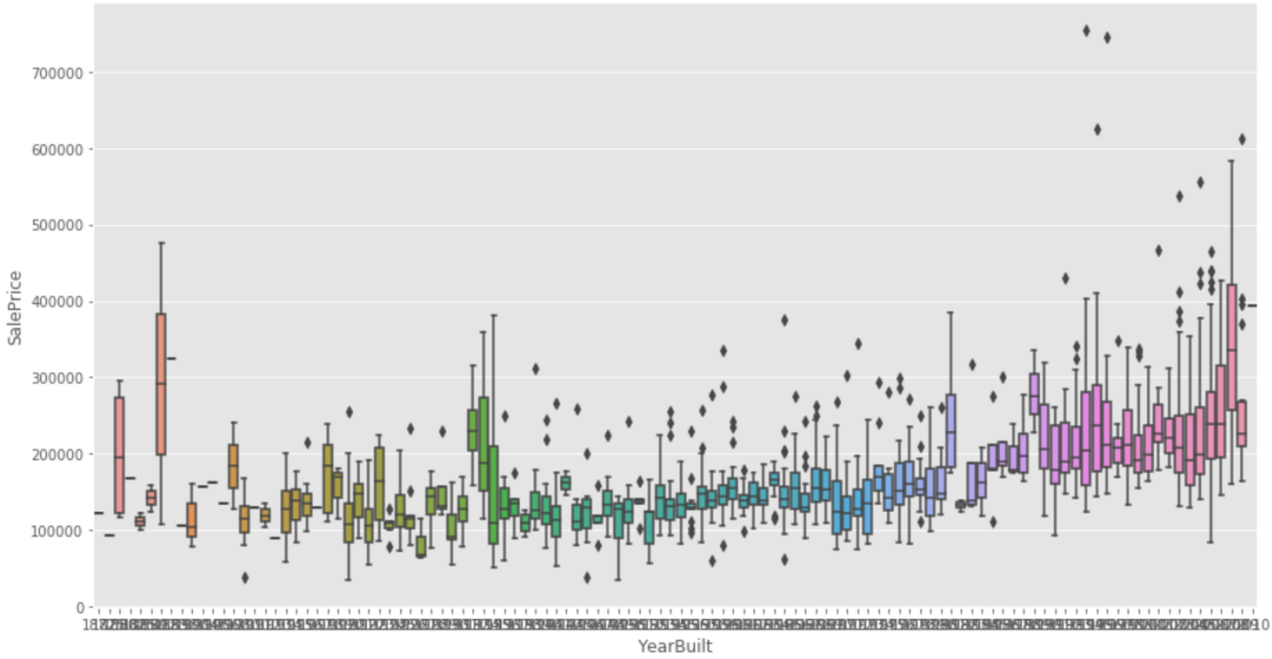


我们首先对预测目标-房屋售价进行探索，线性的模型需要正态分布的目标值才能发挥最大的作用，而在检测房价时候发现其偏离正态分布。此时正态分布明显属于右态分布，整体峰值向左偏离，并且 skewness 较大，需要对目标值做 log 转换，以恢复目标值的正态性。

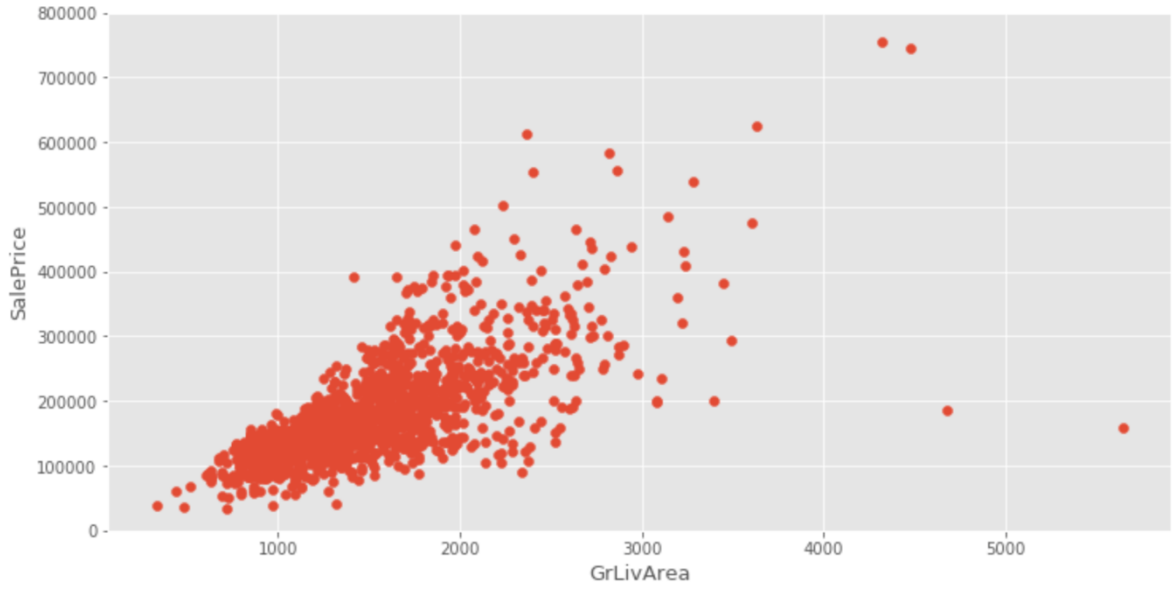


接下来仔细阅读各特征变量所表示的含义，我们发现，有些特征是数值型变量，而另一些特征是类别型变量。

进一步，结合生活经验和业务实际，我们发现一些特征取值变化显著，对房价有显著影响，例如 YearBuilt 建造年份，一般而言，建造年限越久远的房屋售价越低，因此我们绘制出房屋售价随年份变化的盒型图如下图所示，发现大体满足经验规律，因此该特征应该是我们重点考虑的特征之一。



此外，一般来说房屋价格与使用面积也是大体成正相关关系，我们绘制散点图验证如下。



这里我们发现有两个显著偏离的点，因此可以考虑在后面的数据清洗过程中去掉这些点。

我们还认为类似于 OverallQual 等用于评价房屋整体或部分的总体评分的指标也会显著影响到房价，但由于有分类变量的存在，我们需要为所有特征设立一个统一的评价标准，对于缺失的数据 NA 而言，我们需要设计一种填充策略来填补缺失值。因此，接下来的工作会围绕以下几点展开：

- 清除掉数据中明显偏离的点，或不具有普遍性的数据点
- 填补 NA 缺失的数据
- 提取影响房价的关键特征

2、数据清洗

（1）异常值的清除

这里主要借助于散点图的绘制，关键特征对房价的影响一般满足一个普遍规律，对于散点图中显著偏离总体走势的点应予以清除。

（2）数据的合并

由于原始数据中训练和测试集是分开的，因此这里需要暂时合并，而后根据索引再重新分开。合并的主要目的是为了方便进一步的数据清洗和特征提取，由于我们希望对于所有的类别变量建立一个统一的标签编码，因此需要避免在测试集中有，而训练集中没有的类别出现，故此处我们暂时合并训练集和测试集。

（3）缺失数据处理

缺失数据一般有以下几种填补策略：

- 对于大量缺失的数据，可直接舍弃该特征
- 使用平均值、中位数、众数代替，但人为引入随机噪声，导致效果较差
- 使用其他变量预测缺失值，但前提要求其他变量于该变量有较强相关性

对于本数据集而言，通过阅读官方提供的数据说明可知，有些数据缺失是由于该房屋不存在该类型特征，因此对于数值类型数据缺失可以用 0 进行填充，对于其他类型的数据缺失则采用众数进行填充。

3、特征提取

（1）类别型特征的处理

有一些特征表示为类别型变量，以数字表示，但是其并不表示相对大小关系，因此可以将其转换为字符串类型变量。

（2）顺序特征编码

对于有一些类别型变量的特征，可以用数字对其进行编码，创建一个函数建立其类别到编码的一一映射。

（3）添加重要特征

例如，人们往往关心房屋第一层和第二层的面积之和，因此，1StFlrSF 和 2NdFlrSF 之和可以表示为一个新的变量，作为一个特征。除此之外还有许多其他特征需要添加，篇幅原因不一一列举。

五、模型建立

在首先进行数据的清洗，缺失值的处理等数据操作后，最终得到 221 个特征以及 1460 条记录。因为特征数目过于繁多，所以需要进行进一步的特征选择。我们应用 SAS 中的自动的特征选择的选项，进行特征选择，代码如下：

```
proc reg data = WORK.HOUSE_PRICE;
    model SalePrice = MSSubClass--SaleCondition_Partial
    /selection=stepwise slstay=0.05 slentry=0.15;
run;
```

在经过 12 步的逐步选择后，我们得到了 10 个特征用于多元回归，如下表：

“逐步选择”的汇总

步	进入的变量	删除的变量	标签	进入的变量数	偏 R 方	模型 R 方	C(p)	F 值	Pr > F
1	OverallQual		OverallQual	1	0.0823	0.0823	47.7056	130.55	<.0001
2	LotArea		LotArea	2	0.0145	0.0968	26.0045	23.33	<.0001
3	Neighborhood_NridgHt		Neighborhood_NridgHt	3	0.0109	0.1076	10.2317	17.70	<.0001
4	BsmtFullBath		BsmtFullBath	4	0.0063	0.1139	1.9695	10.28	0.0014
5	Electrical_SBrkr		Electrical_SBrkr	5	0.0045	0.1184	-3.4075	7.43	0.0065
6	GrLivArea		GrLivArea	6	0.0041	0.1226	-8.1898	6.85	0.0089
7	YrSold		YrSold	7	0.0039	0.1264	-12.559	6.46	0.0111
8	Exterior1st_BrkFace		Exterior1st_BrkFace	8	0.0039	0.1304	-17.009	6.57	0.0105
9	Exterior2nd_Stucco		Exterior2nd_Stucco	9	0.0034	0.1337	-20.493	5.60	0.0181
10	BsmtQual		BsmtQual	10	0.0025	0.1363	-22.645	4.25	0.0394

对选择出来的 10 个特征进行多元回归，最后结果如下：

1.回归方程的显著性检验

方差分析

源	自由度	平方和	均方	F 值	Pr > F
模型	10	1.254491E12	1.254491E11	22.83	<.0001
误差	1447	7.950795E12	5494675142		
校正合计	1457	9.205286E12			

由该方差分析表可知，P 值小于 0.0001，即该回归模型的因变量与所有的自变量之间存在着一个显著的线性关系，可以使用该模型进行多元回归

2．回归系数的显著性检验

对于每一个自变量来说，他们的 P 值均小于 0.05，这就说明每一个单独的自变量对因变量都有显著的影响，每一个自变量都可以用来进行多元回归。

参数估计						
变量	标签	自由度	参数估计	标准误差	t 值	Pr > |t|
Intercept	Intercept	1	-128564	53581	-2.40	0.0165
OverallQual	OverallQual	1	6792.30567	2157.55542	3.15	0.0017
LotArea	LotArea	1	13953	4145.58310	3.37	0.0008
Neighborhood_NridgHt	Neighborhood_NridgHt	1	38660	9547.50098	4.05	<.0001
BsmtFullBath	BsmtFullBath	1	18230	5741.19788	3.18	0.0015
Electrical_SBrkr	Electrical_SBrkr	1	24381	7316.57451	3.33	0.0009
GrLivArea	GrLivArea	1	18474	8000.75914	2.31	0.0211
YrSold	YrSold	1	-3693.73795	1469.28017	-2.51	0.0120
Exterior1st_BrkFace	Exterior1st_BrkFace	1	29905	10836	2.76	0.0059
Exterior2nd_Stucco	Exterior2nd_Stucco	1	37602	15119	2.49	0.0130
BsmtQual	BsmtQual	1	-4031.27557	1955.20996	-2.06	0.0394

3.多重共线性处理

由于我们已经使用 SAS 中的逐步回归(stepwise regression)将向前选择和向后剔除两种方法结合起来筛选自变量，因此我们使用的这 10 个自变量已经是比较合适的自变量了，多重共线性问题已经解决。

4.拟合优度检验

均方根误差	74126	R 方	0.1363
因变量均值	180971	调整 R 方	0.1303
变异系数	40.96026		

根据拟合优度的表我们可以得到调整 R 方为 0.1303，这也就说明该回归模型中自变量的 13.03%可以有这 10 个自变量解释。鉴于数据的繁杂性，特征的多样性，该结果也是可以接受的。

5.最终模型建立

根据以上的分析，我们确定建立的模型是显著的，每一个自变量是显著的，因此我们可以建立多元回归模型:

$$SalePrice = 6792.3 * OverallQual + 13952 * LotArea + 38660 * Nei + 18230 * Bsmt + 24381 * Ele + 18474 * Gr - 3693.7 * Yr + 29905 * Ex1 + 37602 * Ex2 - 4031.3 * BsmtQual - 128564$$

六、模型检验

经过数据处理与模型建立，我们获得了可用于预测房屋最终售价的多元线性回归模型，该模型是基于 OverallQual、LotArea、Neighborhood_Nridght、BsmtFullBath、Electrical_SBrkr、GrLivArea、YrSold、Exterior1st_BrkFace、Exterior2nd_Stucco 和 BsmtQual 这十个特征变量进行的多元线性回归。为了检验该模型的准确率，我们从测试集中随机选取 200 例数据作为模型的测试对象构建测试数据集 test。然后使用上一节中建立的多元线性回归模型，计算预测的房屋售价 PriceResult，并与测试集数据中给出的实际售价 SalePrice 进行比对，计算模型预测的准确率。准确率的计算公式如下：

$$Accuracy = 1 - \left|\frac{PredicetedPrice - RealPrice}{RealPrice}\right|$$

该过程的代码如下所示：

```
data test;
set work.houseprice(obs = 200);
PriceResult = 6792.3 * OverallQual + 13925 * LotArea + 38660 * neighborhood_nridght
+ 18230 * BsmtFullBath + 24381 * Electrical_SBrkr + 18474 * GrLivArea - 3693.7 * YrSold
+ 29905 * Exterior1st_BrkFace + 37602 * Exterior2nd_Stucco - 4031.3 * BsmtQual - 128564;
Accuracy = 1 - abs(priceResult - saleprice) / saleprice;
run;
```

运行代码可得多元线性回归模型预测的房屋价格，以及该预测价格与实际价格间的比较和准确率。该部分运行结果截图如下，完整结果可见附件资料 result/accuracy_House_Price-results.pdf 文件。

Obs	priceResult	SalePrice	Accuracy
1	197203.13	208500	0.94582
2	177636.56	181500	0.97871
3	201990.96	223500	0.90376
4	198307.24	140000	0.58352
5	215917.18	250000	0.86367
6	182867.89	143000	0.72120
7	218039.61	307000	0.71023
8	200081.61	200000	0.99959
9	178215.23	129900	0.62806
10	165211.64	118000	0.59990
11	170299.20	129500	0.68495
12	275357.74	345000	0.79814
13	169916.22	144000	0.82003

该结果列表中 PriceResult 是多元线性回归模型所预测的结果，SalePrice 是该数据样例对应房屋的实际售价，Accuracy 则是依据本节上文中的公式计算得出的准确率。然后使用 Proc Means 对变量 Accuracy 进行分析，得到如下结果：

The MEANS Procedure

Analysis Variable : Accuracy				
N	Mean	Std Dev	Minimum	Maximum
200	0.7600020	0.2997451	-2.5323420	0.9998598

由表中数据可知，对于样本容量为 200 的测试集数据，该多元线性回归模型对于房屋价格的平均预测准确率可以达到 76%，该预测准确率标准差约为 0.2997。这一准确率对于结构直观且较为朴素的多元线性回归模型是的十分合理的结果，可以得出我们所建立的模型具有完备的功能性与一定的正确性。除此之外，尽管本课程并不要求将模型提交至 Kaggle 的竞赛官网上，我们仍然尝试将最终的回归模型提交至线上竞赛，取得了不错的成绩，详情可见本报告附录中的竞赛结果提交截图。

七、总结

本项目使用 SAS 统计分析语言建立了一个基于选定特征的多元线性房价预测模型。项目数据来源于 Kaggle 竞赛官网[1]，本组对原始数据集进行数据探索后，对数据进行异常值清除、数据合并与数据缺失处理。然后划分训练集与测试集，针对训练集选定十个最优特征值进行多元线性回归，得到房价预测模型的公式。基于该多元线性回归模型，使用测试集对其进行验证，计算房价预测准确率，最终得到平均准确率达到 76%。

该项目是本组作为统计分析与商务智能课程的结课项目，是一学期学习成果的总结与展示。在完成项目的过程中，小组成员们回顾了本学期所学的统计分析的理论知识以及 SAS 程序的编写方法，努力做到活学活用、学以致用。小组成员的专业都是计算机科学技术，虽然编程对我们来说是轻车熟路，但统计分析的理论知识又是一个全新的领域，而 SAS 编程语言便是将二者进行了有机结合。我们在这一过程中不断学习，了解到诸多自身专业知识外的理论，并学会了如何运用这一知识去分析问题、解决问题，成功建立了房屋价格预测模型。可以说对于本组的每一位成员，无论是一学期的学习还是实现项目任务的过程，都是拓宽学习思路、收获颇丰的一段宝贵的经历。

六、小组成员分工

成员	分工
陈麒先	负责数据收集、数据探索、数据清洗以及特征提取； 参与模型设计与验证，得出并整合预测结果，在 kaggle 上完成提交； 撰写报告一至四章内容，进行全文审校与排版。
赵浩森	负责建立模型，分析模型运行结果； 参与数据特征选取，得出最终用于预测的特征； 撰写报告第五章内容。
熊家欣	负责验证模型预测结果，总结模型预测成果； 参与模型设计，并选用 PCA 分析数据集特征； 撰写报告第六、七章内容，进行全文校对； 制作汇报 PPT。

七、参考资料

[1] kaggle 入门实例-预测房价

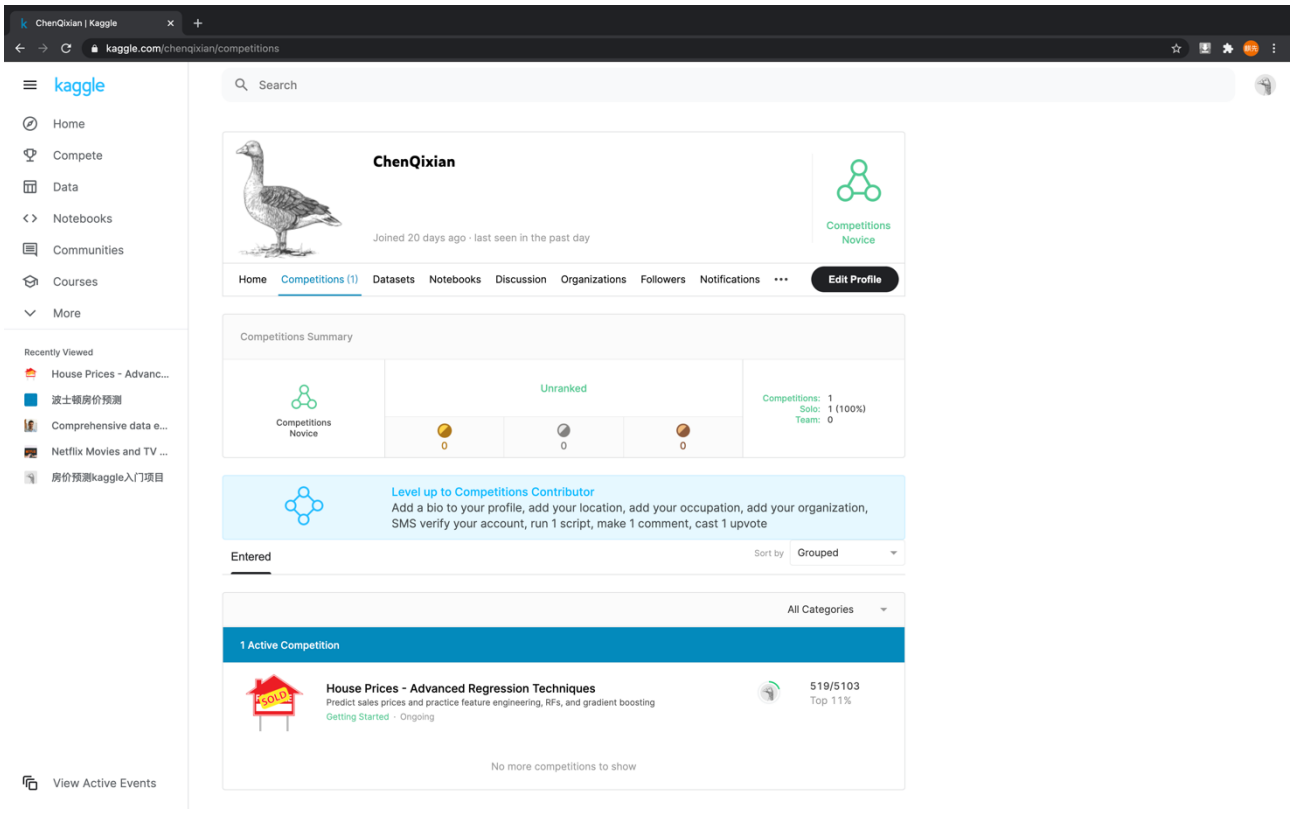
[2] COMPREHENSIVE DATA EXPLORATION WITH PYTHON

[3] Kaggle 竞赛-房价预测（House Prices）小结

[4] 深度学习基础实战 Kaggle 比赛：房价预测

[5] massquantity/Kaggle-HousePrices

附录、竞赛提交结果



*注：以上结果为加入了模型融合的结果，因此成绩较好。最终结果文件由本组陈麒先同学使用账号 ChenQixian 进行提交，真实性可以得到验证。