北华大学
Beihua University

西班牙葡萄酒价格预测

-----C42_2022校赛C题分享

北华大学数学与统计学院

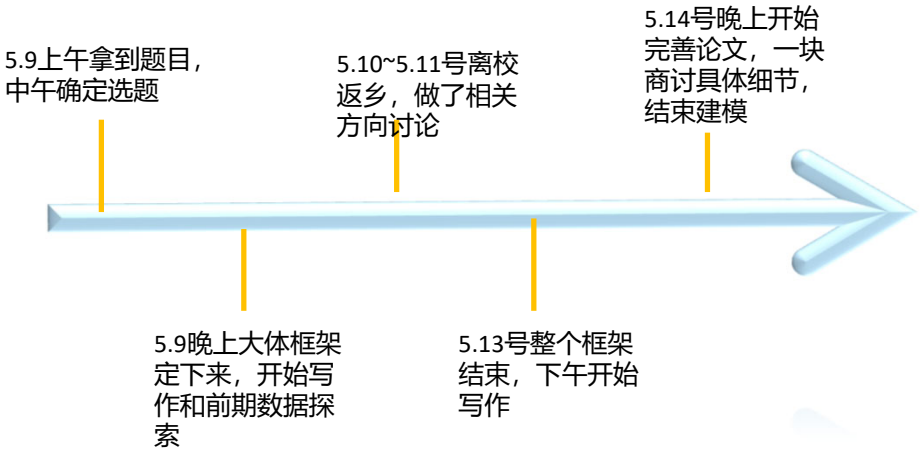
理信19.2 卢　佳

理信19.2 郑国正

理信19.2 陶李涛

2022.06.30

团队分工



团队分工

论文写作、建模过程

建模、代码

代码、画图、写作

问题分析

1. 对数据集进行基础分析，包括但不限于缺失值处理、各变量的统计指标分析、变量间的相关性分析等；

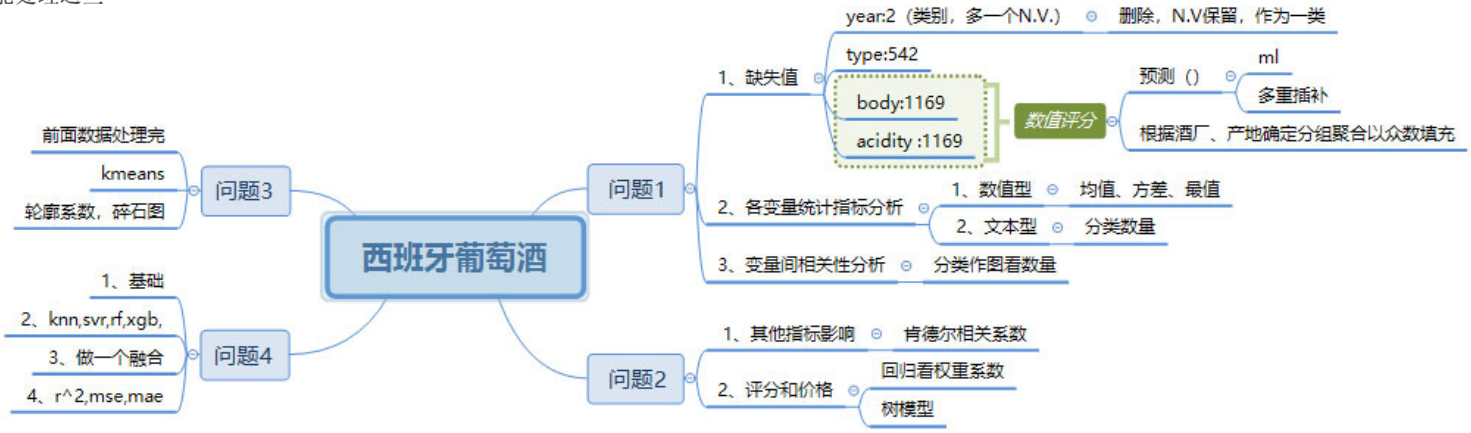
2. 在本数据集中较为重要的是 4 个指标，分别为 Acidity[酸度评分]、body[酒体评分]、rating[用户评分]以及 price[销售价格]。试说明其他指标是如何影响三种评分的，同时阐述评分和价格之间的关系。

3. 对 7500 种葡萄酒进行分类，并详细阐述分类的依据、分类方法和评价分类效果。

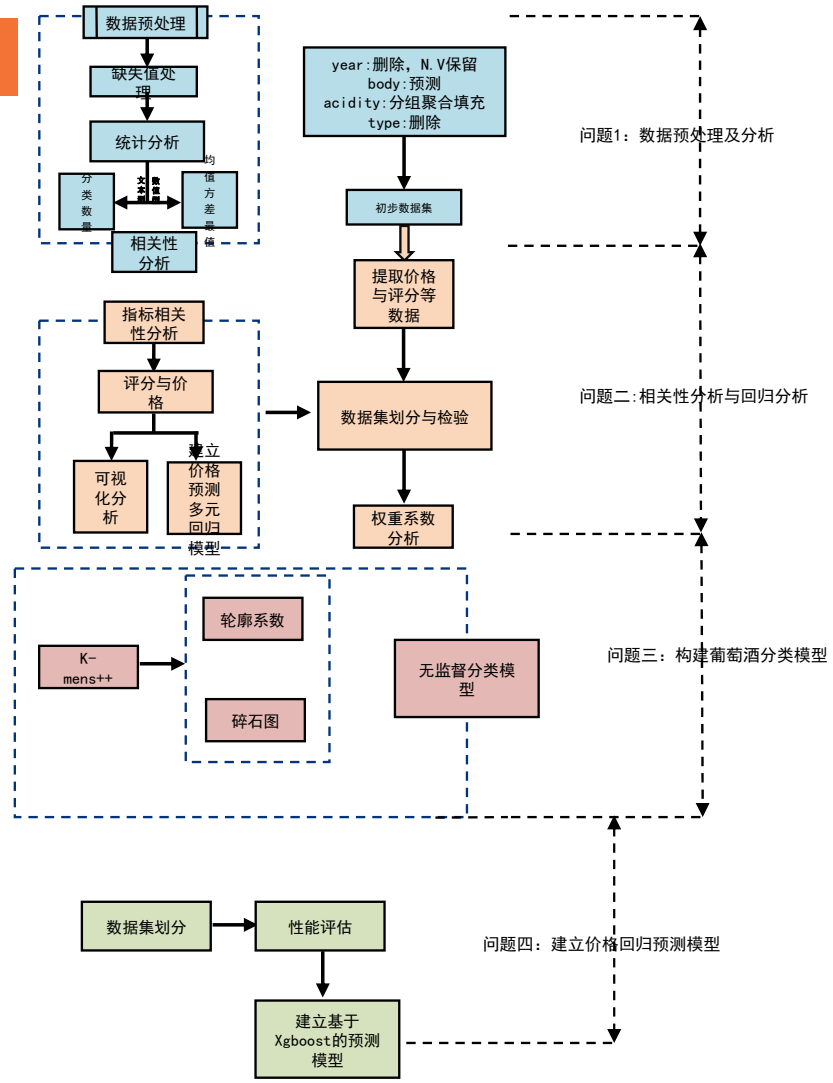
4. 建立葡萄酒的价格预测模型，并利用你的模型确定数据集中编号为 7001-7500 葡萄酒的价格。（需要注意新上架的葡萄酒某些指标会缺失或不可信，如评价人数不可信或评分缺失，你的模型要能处理这些情况）。

index：葡萄酒编号
winery: 酒厂的名字 （文本）
wine: 葡萄酒的名称 （文本）
year: 葡萄收获的年份 （日期）
rating: 用户对葡萄酒的平均评分[评分为1-5分]
num_reviews: 评论葡萄酒的用户数量
region: 葡萄酒的产地 （文本）
type: 葡萄酒品种 （文本）
body: 酒体评分，定义为葡萄酒在口中的丰富度和重量[评分为1-5分]
acidity: 酸度评分，定义为葡萄酒的“皱口”或酸味[评分为1-5分]
price: 价格，单位欧元[€]

葡萄酒酒标上的“NV”是单词“Non-Vintage”的缩写，表示“无年份”的意思，这意味着该款酒并不是由某个单一年份所采摘的葡萄酿成的，而是由不同年份的葡萄酒混合调配而成，这样的标志一般出现在香槟和其他起泡酒上。



建模过程



针对问题1，通过**Python**处理7500种葡萄酒的价格数据，从中查找出**year、type、body和acidity**这4个指标含有缺失值，并对缺失值进行处理；根据均值、方差、最值和四分位数进行统计性描述；利用Python进行数值型变量间的相关性分析，得出多数数值型变量在价格上没有太多的相关性，除了用户评分有一个弱到中等的正相关性.对于类别性变量，采用词云图的方式分析.

针对问题2，运用**肯德尔**系数确定其他指标对三种评分的影响，结合气泡图与箱图分析用户评分和各变量之间的关系；绘制气泡图反映评分与价格的相关性，结合SPSS建立回归方程确定评分与价格的关系，得出**rating、acidity和body**自变量的系数分别为**721.08、12.34和18.27,** 所以rating对价格的影响更显著.

针对问题3，基于K-means聚类将7000种葡萄酒进行分类.观察不同k值对应的簇内误差平方和，绘制**碎石图**得出最优k值等于4；通过轮廓系数对k值进行评估检验，结合散点矩阵图得出主要变量之间具有一定的独立性；最终绘制饼图刻画分组聚类后各个组的占比情况，其中占比最大组为**94.3%,** 占比最小组为**0.1%.**

针对问题4，构建基于Xgboost葡萄酒价格预测模型. 选取mse作为评价指标，按照7：3将数据集进行，使用**autogluon**工具包，选择**KNN、Catboost、Lightgbm**等多个机器学习模型，结果显示**Xgboost**的效果最好，其平均绝对误差为**-9.5133**，根据奥卡姆剃刀原则，选用效果最好的极限梯度提升树来预测葡萄酒价格.

问题建模

问题1：对数据进行可视化呈现分析

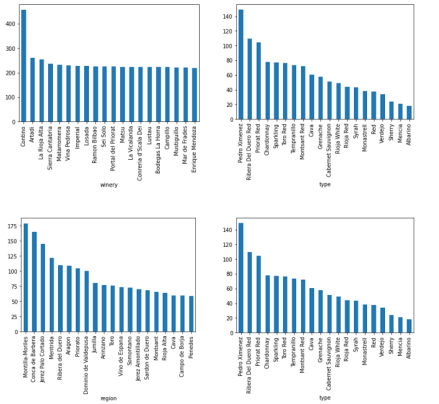


图 4 排名前 20 位的酒厂、产地和葡萄酒品种图

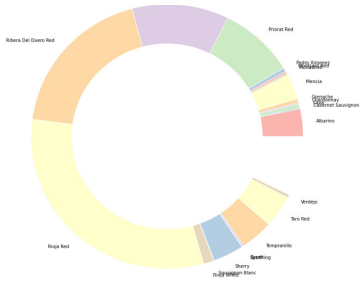


图 5 type 指标中各类别占比情况

图 6 类别型变量统计指标分析

模型假设

(1) 假设原始数据基本准确（个别缺失值数据可处理）；

(2) 假设year指标中N.V.为一个时间类别

(3) 假设数据处理时，对缺失值填充建立的模型对于完整的样本是正确的；

问题建模 问题1：对数据进行可视化呈现分析

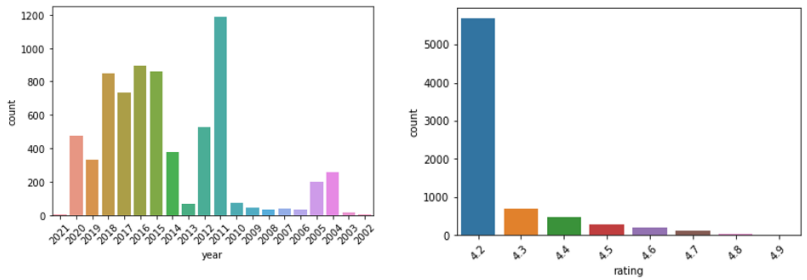


图 7 year 和 rating 指标下各类别出现的频率分布

图 8 各变量间的相关性分析

表 1 各指标的统计性描述

	Rating	Num_eviews	Body	Acidity	Price
count	7500.00	7500.00	6331.00	6331.00	7000.00
mean	4.25	451.11	4.16	2.95	inf
std	0.12	723.00	0.58	0.25	inf
min	4.20	25.00	2.00	1.00	4.99
25%	4.20	389.00	4.00	3.00	19.22
50%	4.20	404.00	4.00	3.00	28.53
75%	4.20	415.00	5.00	3.00	55.00
max	4.90	32624.00	5.00	3.00	3120.00

问题建模

问题2：对变量间进行相关性分析和回归分析

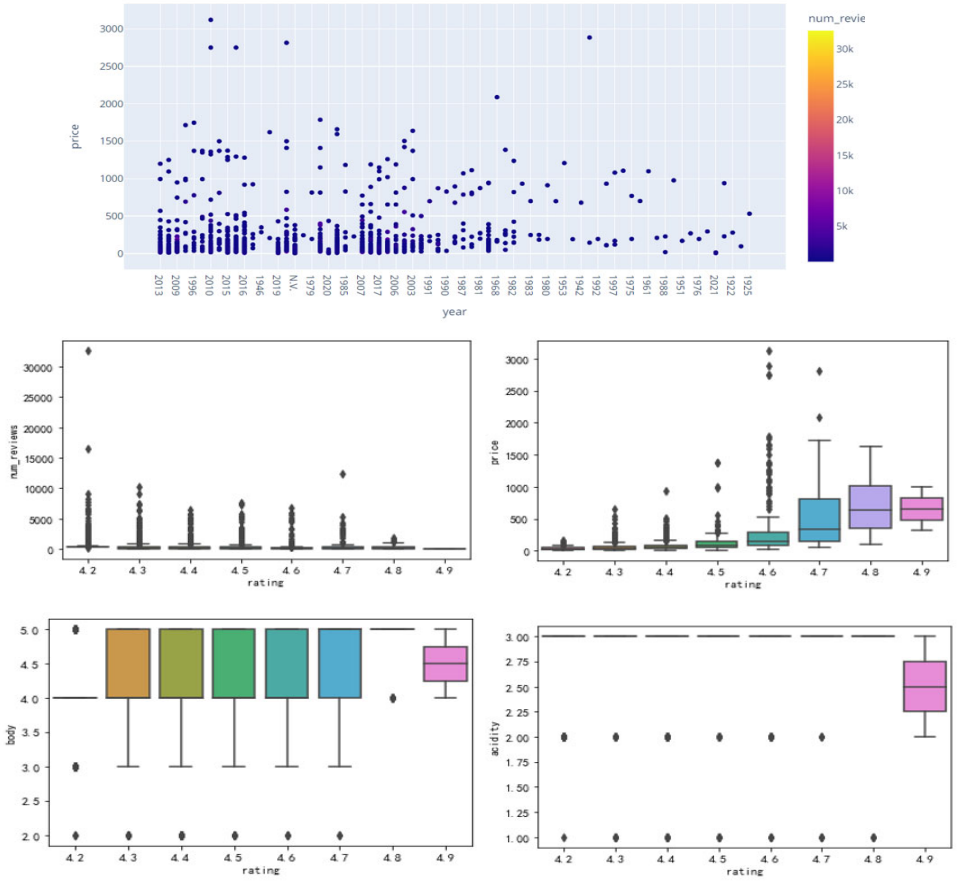
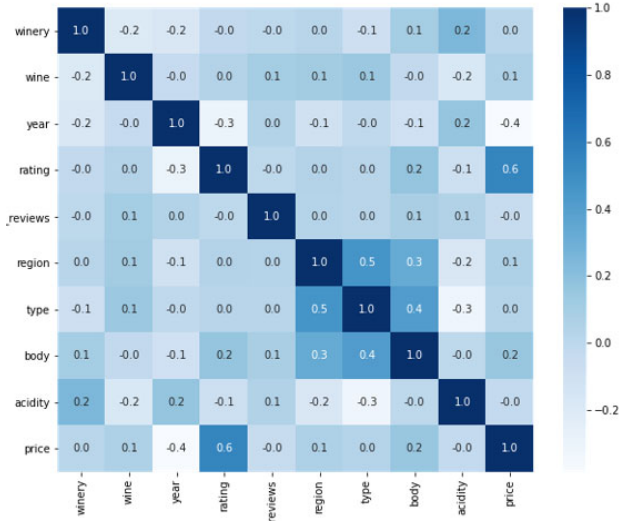


图 10 num_review、price、body、acidity 与 rating 的关系

相关性分析和回归分析



变量	B	标准错误	Beta	t	显著性	容差	VIF
常数项	-3117.67	69.077	0	-45.133	0.000	0	0
rating	721.08	15.01	0.542	48.043	0.000	0.996	1.036
acidity	12.34	7.76	0.018	1.590	0.112	0.992	1.008
body	18.27	3.24	0.063	5.642	0.000	0.973	1.028

操作框架

问题3：进行Kmens++聚类并可视化呈现

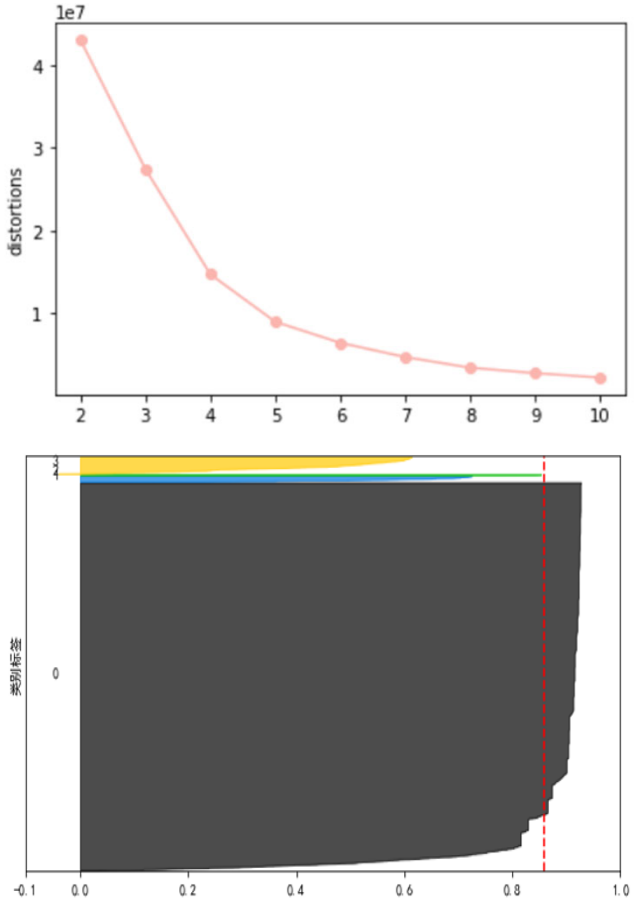


图 14 聚类数目取为 4 时轮廓系数图

	类0	类1	类2	类3
count	5331	82	233	6
mean	inf	inf	inf	2734
std	26.98438	inf	inf	inf
min	6.261719	719.5	170	2088
25%	19.98438	899.5	205	2750
50%	28.95313	1097	259	2782
75%	51.34375	1343.75	350	2866.5
max	170.5	1786	701	3120

任务流程

问题3：进行Kmens++聚类并可视化呈现

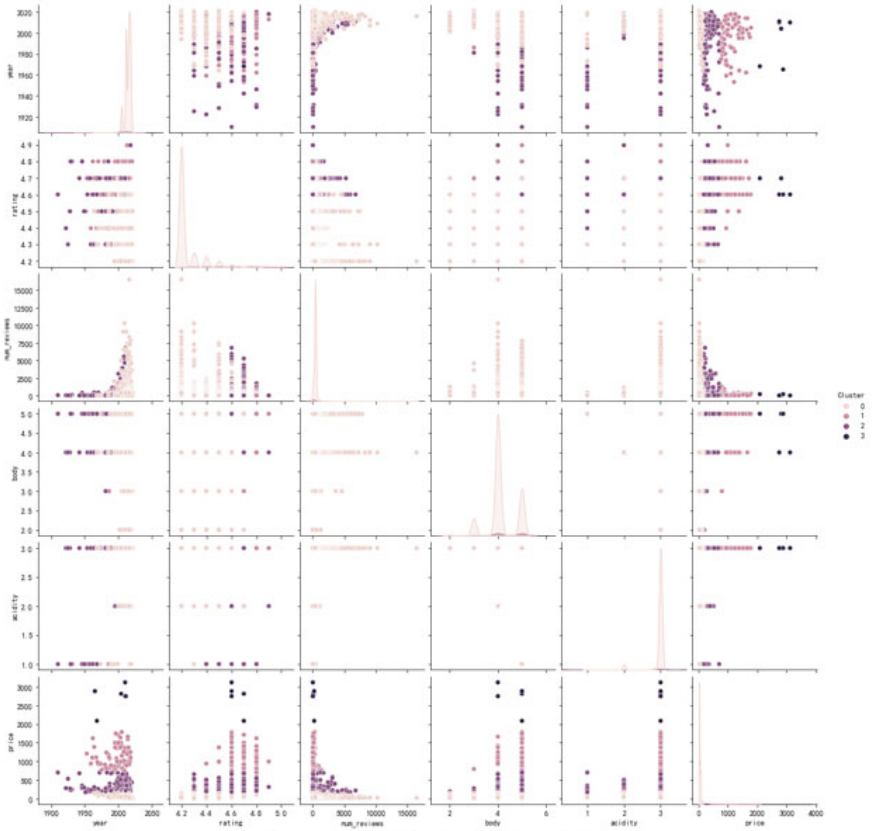


图 15 各变量间的散点图矩阵

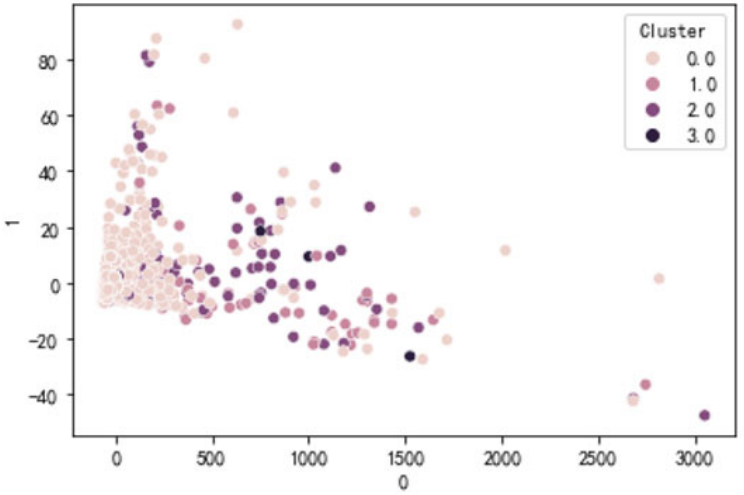


图 16 以 Cluster 为类别制作散点图

任务流程

问题4：使用xgboost模型进行预测价格

不同模型预测结果

model	score_test	score_val
XGBoost	-17.7934	-10.1904
NeuralNetFastAI	-17.8424	-12.613
NeuralNetTorch	-17.9003	-11.513
LightGBMLarge	-18.3767	-11.8541
LightGBMXT	-18.746	-11.8084
LightGBM	-18.9324	-12.0816
RandomForestMSE	-20.0122	-12.9472
CatBoost	-20.1193	-13.5414
ExtraTreesMSE	-20.3221	-13.3837
KNeighborsUnif	-54.9676	-48.1322
KNeighborsDist	-55.3286	-47.1297

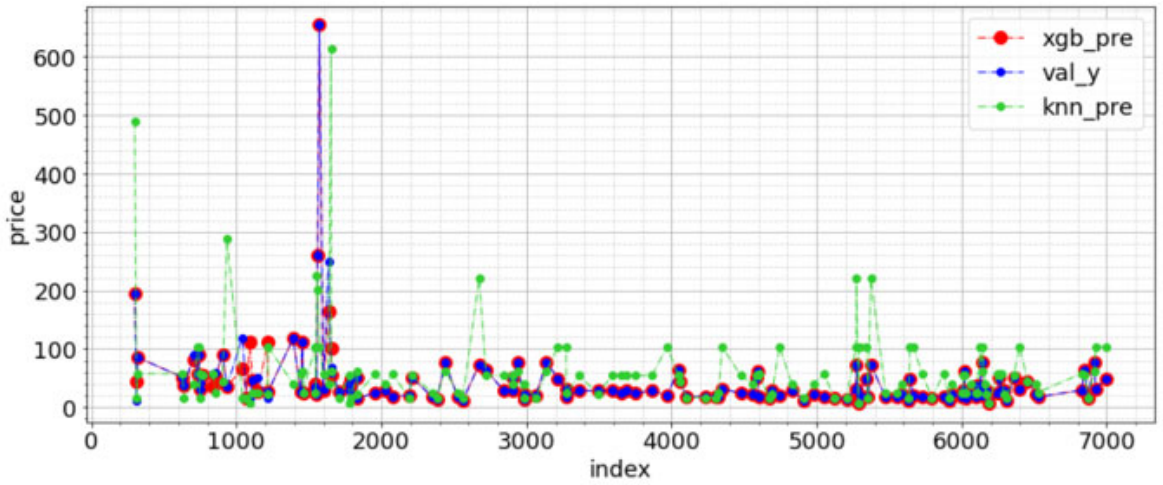


图 18 预测值与真实值的差值

不足与改进

- **one** 模型不够细，需要改的点较多。
- **two** 写作很重要，需要细心，小心!
- **three** 注重检查。
- **four** 团队合作！！！

1.2 问题的提出

在模型的建立过程中，主要解决的问题有 4 个.

问题 1 是对数据集进行基础分析.其中包括缺失值处理、各变量的统计指标分析和变量间的相关性分析.

基于问题 1 中处理后的数据，给出其他指标对三种评分（Acidity、body、rating）的影响并且阐述评分和价格之间的关系.

利用 K-means 算法对 7500 种葡萄酒进行分类.阐述分类的依据、分类方法以及通过轮廓系数与碎石图来评价与分析聚类的效果.

建立葡萄酒的价格预测模型，根据可决系数 R^2、MSE、MAE 来衡量预测模型的准确性，并利用模型确定数据集中编号为 7001-7500 葡萄酒的价格.

参考文献

[1] 刘拥民，罗皓懿，谢铁强. 基于 XGBoost-ARIMA 方法的 PM_(2.5)质量浓度预测模型的研究及应用[J/OL].安全与环境学报:1-13[2022-05-13].DOI:10.13637/j.issn.1009-6094.2022.1849.

[2] 孙林，刘梦含，徐久成. 基于优化初始聚类中心和轮廓系数的 K-means 聚类算法[J].模糊系统与数学,2022,36(01):47-65.

[3] 夏雪，盖靖元. 基于 K-Means 聚类算法的城市轨道交通站点分类及客流特征分析[J].现代城市轨道交通,2021(04):112-118.

[4] 罗春芳，张国华，刘德华，朱定欢.基于 Kmeans 聚类的 XGBoost 集成算法研究[J].计算机时代,2020(10):12-14.DOI:10.16644/j.cnki.cn33-1094/tp.2020.10.004.

[5] 王玉霞，李果，王芳，陈世雄. 基于多元统计分析的葡萄酒及其理化指标评价研究[J].物流工程与管理,2014,36(01):160-164.

建模经验

经验

St1：提交结果的完整性

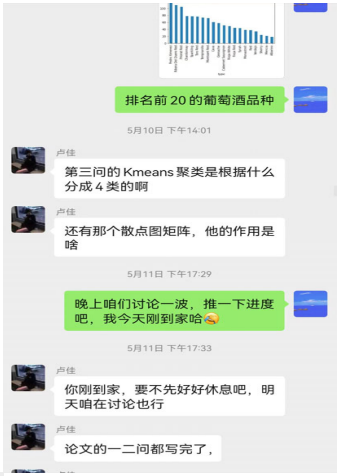
St2：解决问题善于运用手中工具

St3：大步前进，快速迭代

model.pkl	5,730	23,776	PKL 文件	2022-05-14 下午 6:03:...
results.csv	958	2,371	Microsoft Excel 逗号...	2022-05-14 下午 6:51:...
spanish-wine.ipynb	1,806,097	5,343,071	Jupyter 源文件	2022-05-14 下午 8:00:...
submit_results.csv	1,383	7,844	Microsoft Excel 逗号...	2022-05-14 下午 6:03:...
type特征数量分类描述.csv	344	666	Microsoft Excel 逗号...	2022-05-14 下午 6:01:...
type特征数量分类描述图.png	200	1,257	PNG 文件	2022-05-14 下午 6:01:...
xgb.ubj	4,064,829	11,045,424	UBJ 文件	2022-05-14 下午 6:03:...
分类结果表.csv	46,818	488,500	Microsoft Excel 逗号...	2022-05-14 下午 6:02:...
技术流程图.pptx	37,989	46,077	Microsoft PowerPoint...	2022-05-13 下午 9:39:...
价格&三个评分图.png	20,767	24,170	PNG 文件	2022-05-09 下午 10:0...
酒厂名称图.png	276,072	277,931	PNG 文件	2022-05-09 下午 9:50:...
聚类结果分析表.csv	589	1,684	Microsoft Excel 逗号...	2022-05-14 下午 6:02:...
轮廓系数图.png	200	1,257	PNG 文件	2022-05-14 下午 6:02:...
年份图片.png	6,904	7,902	PNG 文件	2022-05-09 下午 9:47:...
评分和价格回归方程系数.png	56,156	57,778	PNG 文件	2022-05-09 下午 10:1...
评分与价格的关系.spv	9,722	10,604	SPSS Statistics Outpu...	2022-05-14 下午 7:52:...
缺失值分布.png	8,348	9,256	PNG 文件	2022-05-14 下午 6:01:...
数值变量皮尔逊相关系数.png	13,790	14,833	PNG 文件	2022-05-09 下午 9:52:...
数值变量相关系数.csv	214	509	Microsoft Excel 逗号...	2022-05-14 下午 6:01:...
数值类型变量描述表.csv	226	406	Microsoft Excel 逗号...	2022-05-14 下午 6:01:...
数值特征表.csv	24,113	171,990	Microsoft Excel 逗号...	2022-05-14 下午 6:01:...
用户评分.png	5,152	5,931	PNG 文件	2022-05-09 下午 9:49:...

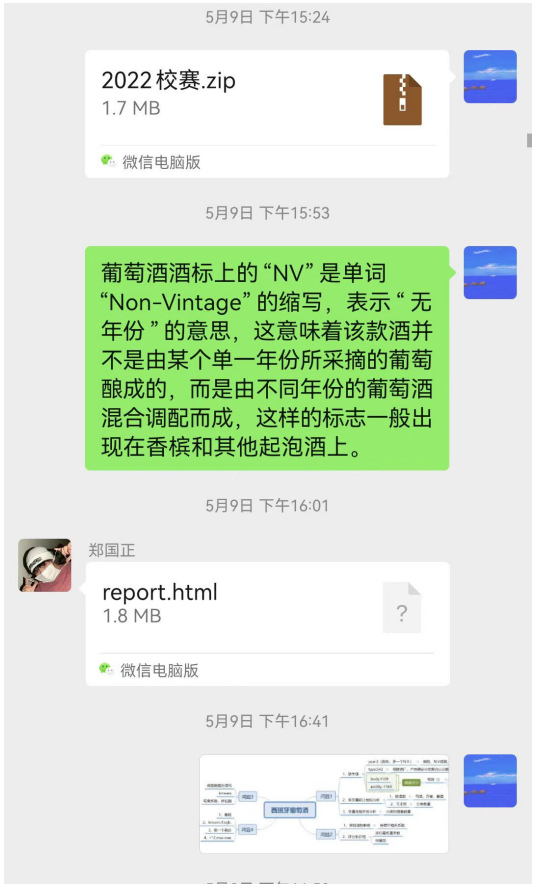
比赛交流


价格整个分布比较偏差比较大
郑国正
是的
主要评价价格看 r^2 吧
结果
5月9日 晚上19:10
我先休息会哈，咱们 8 点讨论一波进展吧
卢佳
你
我
达成共识
5月9日 晚上20:50
咱们继续哈


排名前 20 的葡萄酒品种
5月10日 下午14:01
卢佳
第三问的 Kmeans 聚类是根据什么分成 4 类的啊
卢佳
还有那个散点图矩阵，他的作用是啥
5月11日 下午17:29
晚上咱们讨论一波，推一下进度吧，我今天刚到家哈
5月11日 下午17:33
卢佳
你刚到家，要不先好好休息吧，明天咱在讨论也行
卢佳
论文的一二问都写完了，


5月13日 下午14:05
卢佳
submit_results.csv
7.7 KB
微信电脑版
卢佳
@ 这个是什么意思啊
卢佳


5月9日 下午16:50
5月9日 下午16:55
我的论文.docx
1.2 MB
微信电脑版
5月9日 晚上18:27
@Z 现在最好的模型跑到多少呀
mse 或者 mae 作为评价指标的话
郑国正
mse 太大了
这是那个指标呢
郑国正
mse


5月9日 下午15:24
2022 校赛.zip
1.7 MB
微信电脑版
5月9日 下午15:53
葡萄酒酒标上的“NV”是单词“Non-Vintage”的缩写，表示“无年份”的意思，这意味着该款酒并不是由某个单一年份所采摘的葡萄酿成的，而是由不同年份的葡萄酒混合调配而成，这样的标志一般出现在香槟和其他起泡酒上。
5月9日 下午16:01
郑国正
report.html
1.8 MB
微信电脑版
5月9日 下午16:41


5月14日 晚上19:49
C42_卢佳_郑国正_陶李涛.docx
1.3 MB
微信电脑版
大家有时间的检查一下，看看格式啥的
5月14日 晚上19:55
页码错了
卢佳
还有图 15
卢佳
他的那个图例
15 和 16 顺序标反了都
卢佳
嗷嗷嗯呢


欢迎交流
理信19.2 陶李涛