

《史记·酈生陆贾列传》曾写到：王以民为本，民以食为天。现今，所谓“食”，已不单单是满足维持生命这一基本诉求，更多的是人们追求享受、助于社交以及饮食健康的多方面层次追求。坊间有言：没有什么事是一顿饭解决不了的，一顿不行，那就两顿！但市场上那吃饭的地儿是千千万，那质量也有很多淮南之枳。有时候吃的不舒服，还掏了大把银子，嘿，那心里给憋屈的，法治社会，一个字儿：忍！但估摸着也有些豪绅手里拽着“刀了”“但没找到符合他品味和身份的理想美食天地，这时候心里也难受。

最近笔者在外饭吃的不少，然而真正吃的开心并不多，感觉损失了一个亿。闲暇之余，想在家坐小板凳上透过小小的屏幕看看这广袤的上海滩在衣食住行的“食”这块，有没有什么好玩的能挖掘挖掘，万一挖出个商机啥的，那就定个小目标，比如先挣他一个亿回回本。

以下本文将从数据预处理、数据分析和展现、机器学习和总结这四个部分对上海的餐饮现状做一个粗略和简要的分析。

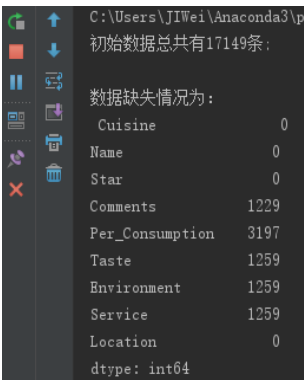
所有代码和分析结果详见：https://github.com/zgjsjw/4th_lesson-SATISFY_YOUR_MOUTH

一、数据预处理

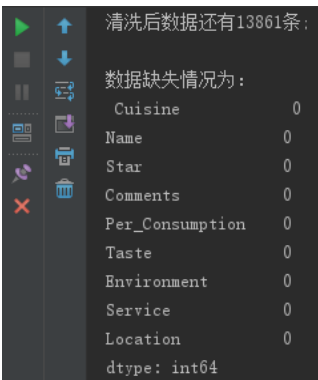
本文所使用的原始数据是通过爬取大众点评网上 1 万+位于上海的餐厅信息，共计 9 个维度:分别包括菜系、餐厅名称、星级、评论数、人均消费、口味评分、环境评分、服务评分和餐厅地址信息，如下事例所示。

	A	B	C	D	E	F	G	H	I	J	K
1	Cuisine	Name	Star	Comments	Per_Consumption	Taste	Environment	Service	Location		
2	蟹宴	七欣天品蟹	45	804	106	8.7	8.3	7.8	文诚路500弄2号102室		
3	蟹宴	大麦live-	45	117	241	7.5	8.3	8.1	北京西路1825号b座		
4	蟹宴	七欣天品蟹	40	494	95	8.5	8	8.2	乐都西路835号		
5	蟹宴	赖胖子肉蟹	40	2529	80	7.7	8	7.9	张杨北路801号文峰广场5层		
6	蟹宴	七欣天迷踪	45	653	97	8.8	8.5	8.4	卫清西路350号		
7	蟹宴	正宜丰	40	424	355	8.3	8.2	7.9	天平路137号		
8	蟹宴	七欣天品蟹	40	1128	145	8.3	7.9	7.8	漕宝路3509号汇宝购物广场A座7		
9	蟹宴	七欣天品蟹	40	1342	89	8.4	7.7	7.8	博乐路85号		

首先对原始数据的缺失值和异常值进行处理。



C:\Users\JIWei\Anaconda3\p
初始数据总共有17149条:
数据缺失情况为:
Cuisine 0
Name 0
Star 0
Comments 1229
Per_Consumption 3197
Taste 1259
Environment 1259
Service 1259
Location 0
dtype: int64



清洗后数据还有13861条:
数据缺失情况为:
Cuisine 0
Name 0
Star 0
Comments 0
Per_Consumption 0
Taste 0
Environment 0
Service 0
Location 0
dtype: int64

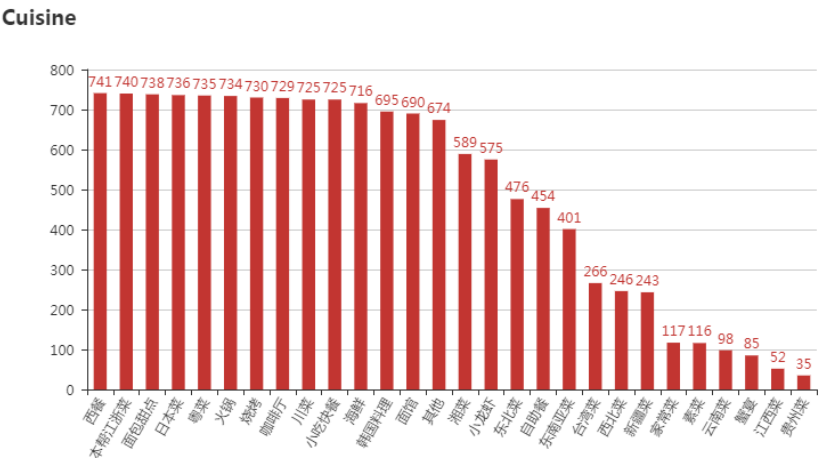
初始数据总共有 17149 条，其中没有菜系、餐厅名称和地址的缺失，但评分项缺失数据较多，由于后续文章需要对评分项和餐厅等级之间的关系进行分析，所以如果对缺失数据进行 0 值填充或者随机填充，那必然会对最后的分析结果造成误差，同时考虑到缺失数据占比较小，故作剔除操作，笔者相信是金子总会发光的。对于重复数据而言，同样做剔除操作，

优秀大家都懂的，要低调。

由于原始数据较为规整，对数据预处理带来了很大便利，所以数据处理部分很容易就可以实现，下面对处理后的数据从各个维度进行简要分析和查看。

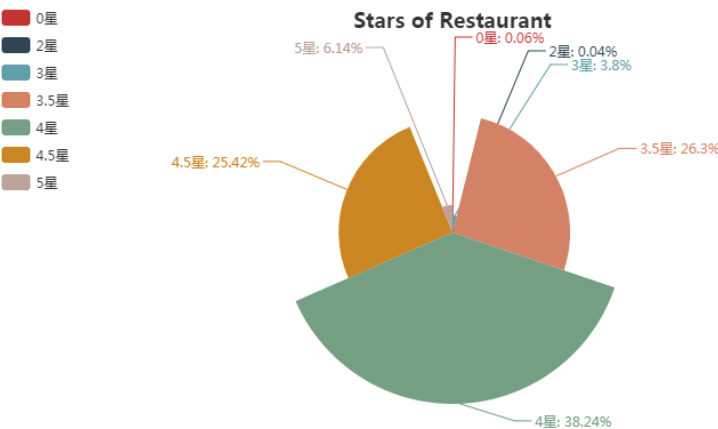
二、数据分析、展现

首先看看上海餐饮的菜系分布情况。



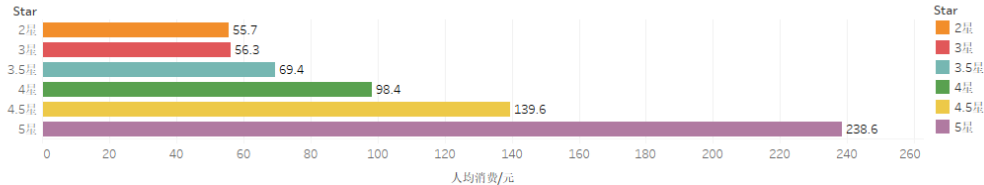
上海作为东方小巴黎，国际化氛围很浓，常住的西方人口也较多，西餐和面包甜点占据了前三甲中冠军和季军的宝座。同时她毗邻江浙，上海本帮菜文化和江浙如出一辙，所以本帮江浙菜数量众多也是情理之中。同时上海作为一个外来常住人口 970 万，户籍常住人口 1450 万的大城市，吸引了天南海北各地的人才汇集于此，大家抱着要想干好活，总得吃饱饭的念头，吃还想着最好能带点家乡的味道，所以也形成了一种多元的饮食文化，粤菜、川菜、东北菜等等在上海也广受欢迎。饮食一定程度上也体现了城市的包容特性，“海纳百川，大气谦和”，这是上海城市精神，所以我们会看到越来越多爱吃甜食的上海人、江浙人，在川菜馆子，在火锅店，辣的一把鼻涕一把泪地大快朵颐。包容开放的城市，我相信她的发展会越来越好，房价会越来越高，房价？哎~心累。。看看人家米国，汇集全球文化，各种思想文化的火花绽放，造就了如今的帝国，听起来好有道理的样子哦~~

接下来，嗯。。我们不算面相，所以餐厅名字就不分析了，来看看上海餐厅的等级情况



从饼状图很容易发现，3.5 星、4 星和 4.5 星占比达到了 90%左右，其中 4 星餐厅几乎快达到了 40 个百分点，说明上海餐饮的整体水平还是很高的，但竞争也很激烈啊，大家都力求做得更好来争取更多的市场份额。用过大众点评的应该都知道，大家在选择餐厅的时候往往更多关注 4 星以上的餐厅，所以店家肯定也会通过各种营销手段将自家星级达到 4 星及以上以获取更多的用户关注。

星级--人均消费分布



从 4 星到 4.5 星需要店家投入一定的成本来达到更好地体验，但消费者肯定也会为此买单，从上图星级和人均消费的分布情况可以看出，4.5 星的人均消费会比 4 星多出 40 元左右，从笔者的消费经历来看，这是相当一部分人群可以接受的范围。但从 4.5 星到 5 星的餐厅数量减少颇多，可以猜测餐厅需要投入更多的成本，相应的消费也会更高，数据显示人均消费在 240 左右，估测对应的消费群体也较少，相应的餐厅数量也就不会很多。

刚才星级分析提到人均消费情况，那接下来就瞅瞅上海餐厅的整体消费水平。

人均最高5238.0元	
人均最低3.0元	
平均消费108.27元	
我（价）最（格）想（最）吃（贵）：	
Cuisine	西餐
Name	Ultraviolet by Paul Pairet
Star	50
Comments	448
Per_Consumption	5238
Taste	9.1
Environment	9.3
Service	9.3
Location	中山东一路18号6楼
Name: 10633, dtype: object	

其中，人均消费最高的是 5238 元，人均最低的是 3 元，平均消费是 108 块 2 毛 7，嗯笔者经历差不多每次出去吃饭人均都在 120 左右，看来是超出了平均水平呀，但貌似收入并没有超啊~悲伤辣么大。。。最高的人均消费是 5000 多，出于好奇就看了下是哪家餐厅如此有格调，别名紫外线餐厅!! 大众点评看了下他家的基本情况。

Ultraviolet by Paul Pairet

手机买单 积分抵现

添加分店

678条评论 人均：5175元 口味：9.2 环境：9.3 服务：9.3

地址：中山东一路18号6楼

电话：无 添加

收起

别名：紫外线

营业时间：周二至周六 19:00-23:30 修改

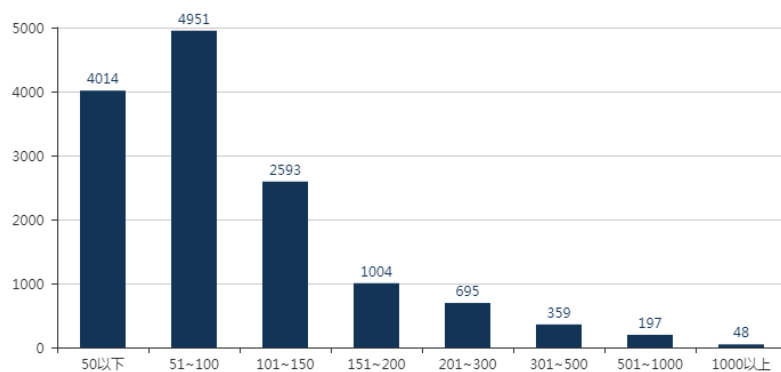
[前卫分子料理 \(245\)](#)
[松露面包 \(126\)](#)
[鹅肝巧克力派 \(101\)](#)
[花雕鱼头锅 \(60\)](#)
[萝卜蛋糕 \(54\)](#)
[笋子鸡 \(48\)](#)
[冷锅耗儿鱼 \(40\)](#)

[lobster \(39\)](#)
[梭边鱼火锅 \(33\)](#)
[老坛酸菜牛肉面 \(30\)](#)
[麻辣鲜香冷锅鱼 \(30\)](#)
[爆炒海菜 \(20\)](#)
[油封松露羊排 \(18\)](#)
[爆炒黑木耳 \(15\)](#)

[更多 >](#)

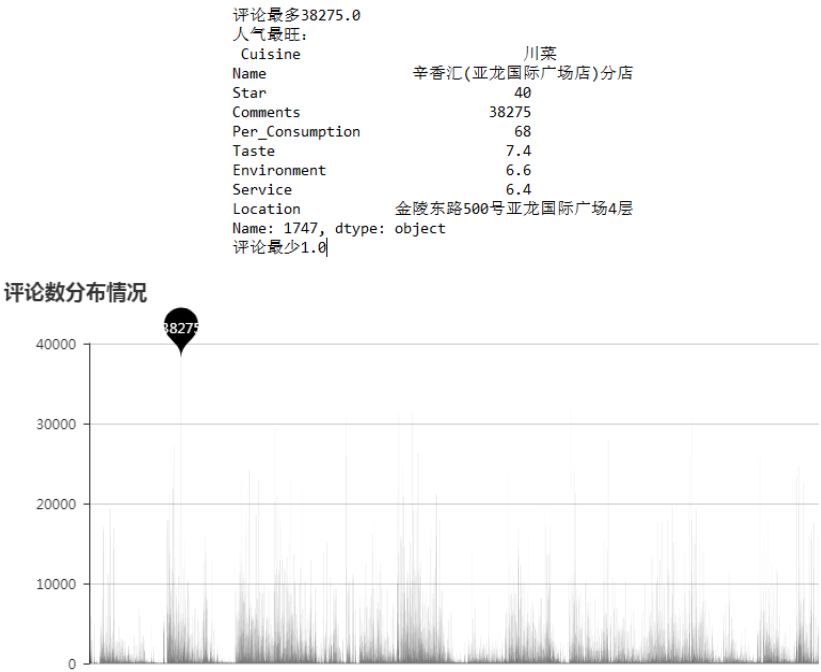


人均消费分布



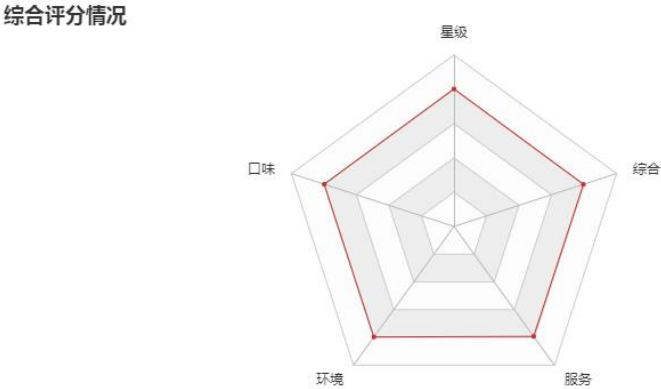
可见，绝大多数人的消费还是在 150 以下，这也是大多数普通工薪阶层的消费水平和能力，但也不乏有些豪绅，消费在 500 至 1000，甚至 1000 以上的水平，我们要向别人家的孩子看齐，努力奋斗，争取下次达到人均消费四位数，争取不到就去睡个觉~其实有 48 家餐厅的人均消费达到了上千的档次，数量还是不少的，看来贫富差距还是比较大的。

看完平均消费情况，找到了一个最贵餐厅，摸摸裤兜，还是决定放弃。那就再来分析下评论数，找找最受欢迎的餐厅，或许能消费得起去试试。



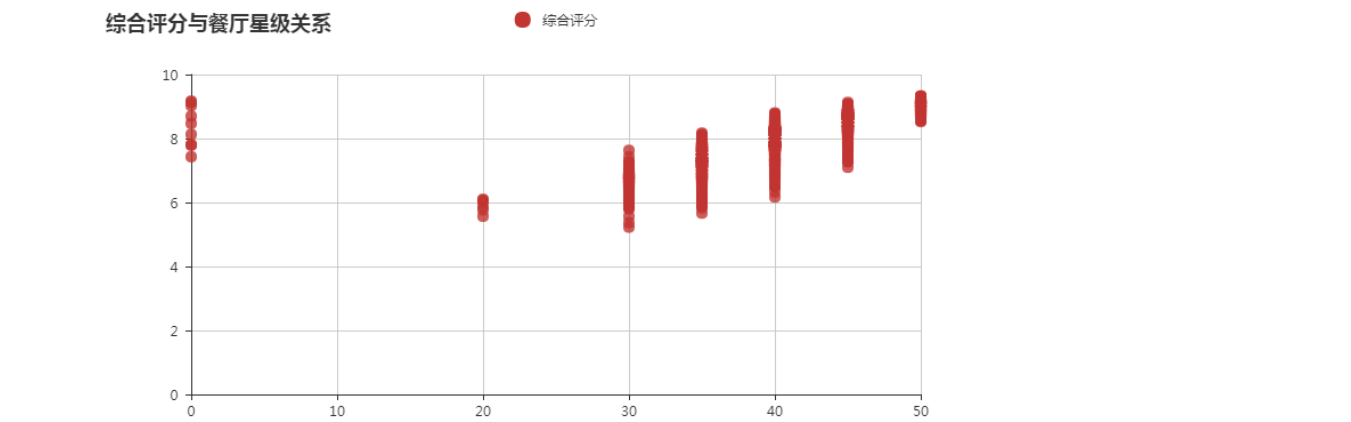
评论数最多的是一家川菜馆，有 38275 条评论，其人均消费也不高，约 70 元左右，这是很多消费者能够接受的范围。从评论数分布情况可以看得出，绝大多数商家的评论数在 1 万条以下，极少数个别商家能有 3 万以上的评论，可能是这些大牌商家比较具有营销手段能够通过各种活动在短时间内吸引大量客流，比如评论、点赞实现减免消费之类的。

取口味、环境、服务的平均得分为综合得分参数指标，分析下上海所有餐厅的一个综合水准分布，从下图可以看出，其各项综合指标分布非常均衡，没有明显的优势和短板，这也应该是一个正常的市场化结果，说明上海整体的餐饮市场水平发展还是比较健康的。



紧接着，来看看综合评分和星级之间的关系情况，同预想的一样，餐厅星级越高，其整

体综合评分也是呈现正相关上升趋势。对 0 星级的数据，可能由于餐厅刚成立，还没有评级结果，但是已有部分评分数据，而且综合评分分数较高，不排除里面隐藏着很多潜力股高质量餐厅，相信只要做得好，一定会得到市场和群众的认可。



最后，从这些数据中找找一些餐厅之最，诸如人均消费最高，最受欢迎（评论数最多），各项评分指标最高的那些餐厅。

最昂贵的菜系：日本菜 人均：236.32元
最受欢迎的菜系：火锅 平均有：3047条评论

Cuisine	Name	Star	Comments	Per Consumption	Taste	Environment	Service	Location
蟹宴	黄记阳澄湖大闸蟹专卖店	50	374	170	9.3	9	9.3	乌鲁木齐中路308号
海鲜	大钊锅物	50	402	337	9.3	9.3	9.3	福山路312号世纪大都会B1层B110B室
贵州菜	多彩贵州	50	140	38	9.3	9	9.3	四平路716号
本帮江浙菜	黄记阳澄湖大闸蟹专卖店	50	374	170	9.3	9	9.3	乌鲁木齐中路308号
火锅	季悦火锅(锦江店)分店	50	446	886	9.3	9.4	9.3	茂名南路59号
火锅	赤鼎麻辣锅(巴黎春天店)	50	621	130	9.3	9.2	9.4	长寿路155号巴黎春天5楼
火锅	大钊锅物	50	402	337	9.3	9.3	9.3	福山路312号世纪大都会B1层B110B室
火锅	季悦火锅(虹桥迎宾馆店)分店	50	466	1012	9.3	9.3	9.3	虹桥路1591号虹桥迎宾馆内6号楼2楼
西餐	Reve Kitchen创意法式料理	50	77	232	9.3	9.3	9.3	申长路988号虹桥万科中心LG229号
西餐	红花铁板烧	50	902	745	9.3	9.1	9.4	古羊路1129号名都城俱乐部后廊
其他	樽宴	50	104	1352	9.3	9.1	9.3	芳甸路777号往北第二个路口右转，第一栋281号
其他	关小刀即时捞	50	103	94	9.3	9	9.3	尚博路687号
粤菜	瑞华樟园广告	50	273	545	9.1	9.4	9.1	凯旋北路1555弄66号
本帮江浙菜	西郊壹号分店	50	607	682	9.1	9.4	9.1	虹桥路1950号
火锅	季悦火锅(锦江店)分店	50	446	886	9.3	9.4	9.3	茂名南路59号
火锅	锅季一捞王旗下品牌(小火锅)	50	1019	126	9.2	9.4	9.4	申长路869号A馆2楼L2-31号
日本菜	ANTHOLOGIA地球美食剧场	50	591	1134	9	9.4	9.3	番禺路381号D栋105-107室
西餐	麒麟阁	45	125	193	8.7	9.4	8.6	金科路东郊宾馆内
西餐	Aurora 绚景楼	50	107	921	9.1	9.4	9.4	上海迪士尼乐园酒店八楼
火锅	赤鼎麻辣锅(巴黎春天店)	50	621	130	9.3	9.2	9.4	长寿路155号巴黎春天5楼
火锅	潮界(绿地缤纷城店)	50	199	75	9.2	9	9.4	东安路562号绿地缤纷城东区h201
火锅	海底捞火锅(龙茗路店)分店	50	2884	126	9.2	9.2	9.4	龙茗路1300弄58号古美生活购物广场三层
火锅	锅季一捞王旗下品牌(小火锅)	50	1019	126	9.2	9.4	9.4	申长路869号A馆2楼L2-31号
西餐	O2ZONE氧舍(中山北路店)	50	587	73	9.2	9.3	9.4	中山北路756号
西餐	PizzaD	50	269	60	9.2	8.7	9.4	东昌路508号
西餐	红花铁板烧	50	902	745	9.3	9.1	9.4	古羊路1129号名都城俱乐部后廊
西餐	Aurora 绚景楼	50	107	921	9.1	9.4	9.4	上海迪士尼乐园酒店八楼

人均消费最高的是日本菜，人均 236.32 元，说明日本料理相对而言还是比较贵的，最受欢迎的菜系是火锅，平均有 3047 条评论数，火锅作为中国独创的美食，历史悠久，成为如今中国广大人民群众休闲聚餐、建立友谊、加深合作、巩固亲情的优选方式也是情理之中。此外，还将口味、环境和服务评分在 9.0 分以上的餐厅抓取了出来，其中一些人均消费也还可以接受，可以作为以后觅食之处的优先之选。

三、机器学习

以上对数据做了一些简要的分析，接下来想利用一些简单常用的机器学习算法对餐厅的

质量做一个分类处理，由于笔者目前自学知识和时间精力有限，先考虑做一个二分类问题。将 4 星及以上的餐厅认为是好餐厅，将 4 星以下的餐厅认为是还需要努力提升的餐厅，因此选择['Star']字段作为 label，选取 feature 如下：

['Cuisine', 'Comments', 'Per_Consumption', 'Taste', 'Environment', 'Service']

餐厅名称和地理位置对于分类没有特别大的相关性，所以不考虑作为特征来处理。

1、首先需要对数据做一些预处理工作：

➤ label 二值化

由于星级数据是几个离散变量，所以需要设定阈值对离散数值进行[0, 1]二值化处理，对 4 星及以上设置标签为 1，对 4 星以下设置标签为 0；



```
Star二值化结果:
C:\Users\JIWei\Anaconda3\lib
[0 1 0 ... 1 1 0]
```

➤ object 类别变量转换数值变量

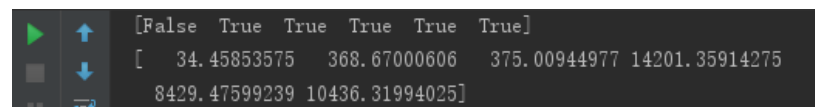
菜系为字符串非数值标签，这里通过 label_encode 进行标签转换，用于后续的特征计算；



```
C:\Users\JIWei\Desktop\WHERE I
菜系变量转数值变量:
[20 16 22 ... 2 23 15]
```

➤ 特征重要性选取

接下来需要对特征进行一个重要性程度排序，选取前 85%的特征作为最终用来训练的和预测的特征

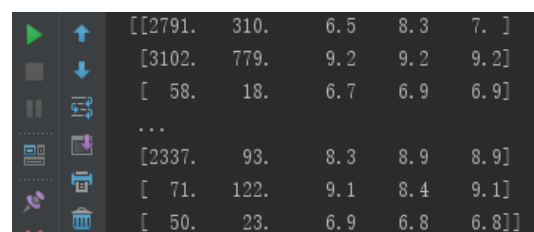


```
[False True True True True True]
[ 34.45853575 368.67000606 375.00944977 14201.35914275
 8429.47599239 10436.31994025]
```

由于评论数特征和消费特征重要性程度很接近，所以设置阈值 85%，去掉菜系特征，用['Comments', 'Per_Consumption', 'Taste', 'Environment', 'Service']这 5 个特征进行计算。

➤ 特征归一化处理

由于评论数和人均消费数值相对评分来看较大，需要对其进行缩放，以防特征被主导影响预测结果。



```
[[2791. 310. 6.5 8.3 7. ]
 [3102. 779. 9.2 9.2 9.2]
 [ 58. 18. 6.7 6.9 6.9]
 ...
 [2337. 93. 8.3 8.9 8.9]
 [ 71. 122. 9.1 8.4 9.1]
 [ 50. 23. 6.9 6.8 6.8]]
```

通过 MinMaxScaler 归一化，进行后续特征计算

```

[[0.07289544 0.06099742 0.31707317 0.76595745 0.45454545]
 [0.08102106 0.1541824 0.97560976 0.95744681 0.95454545]
 [0.00148926 0.00298033 0.36585366 0.46808511 0.43181818]
 ...
 [0.0610336 0.01788198 0.75609756 0.89361702 0.88636364]
 [0.00182892 0.02364395 0.95121951 0.78723404 0.93181818]
 [0.00128024 0.00397377 0.41463415 0.44680851 0.40909091]]

```

2、以 4:1 的数据维度将数据集拆分成训练集和测试集，选取常用的 kNN、LR、DT、SVM 这四种分类方法对训练集数据进行交叉验证，选取 `cv=5`，利用测试集作为准确率验证数据，计算准确率和时长作为算法评价指标。其中算法的初步参数选择如下

```

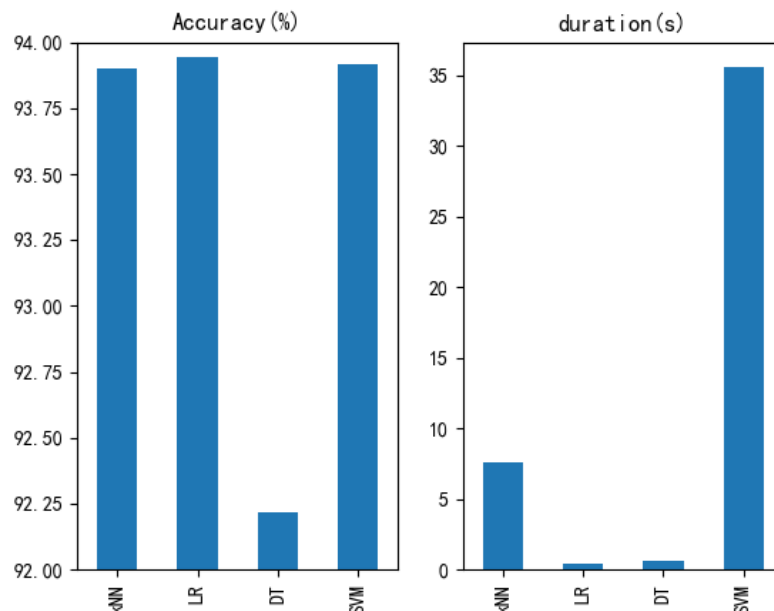
# 数据建模和验证
print('\n-----数据建模及验证-----\n')
model_para_dic = {'kNN': (KNeighborsClassifier(), {'n_neighbors': [5, 20, 50]}),
                  'LR': (LogisticRegression(), {'C': [0.01, 1, 100]}),
                  'DT': (DecisionTreeClassifier(), {'max_depth': [10, 30, 80]}),
                  'SVM': (SVC(), {'C': [0.01, 1, 100]})}
result = pd.DataFrame(columns=['Accuracy(%)', 'duration(s)'], index=model_para_dic.keys())

```

通过交叉验证，得到模型测试准确率和训练耗时如下

	Accuracy(%)	duration(s)
kNN	93.901	7.6095
LR	93.942	0.4134
DT	92.22	0.6035
SVM	93.914	35.5691

将性能指标绘制成柱状图，可以更直观一些



3、从上图可以发现，LR 分类方法表现最好，因此选取 LR 作为分类预测算法，进行参数调优，尝试进一步提升预测准确率。


```

    paral = {'C': [i/100 for i in (1, 10000)],
             'intercept_scaling': [i/10 for i in (1, 100)],
             'max_iter': [i for i in (1, 200)],
             'tol': [i/10000 for i in (1, 100)]}

    lr1 = GridSearchCV(estimator=
        LogisticRegression(C=1.0, class_weight=None, dual=False, fit_intercept=True,
                           intercept_scaling=1, max_iter=100, multi_class='ovr', n_jobs=1,
                           penalty='l2', random_state=None, solver='liblinear', tol=0.0001,
                           verbose=0, warm_start=False),
        param_grid=paral,
        scoring='f1',
        cv=5,
        refit=True)

```

```

data['Cuisine'] = pre.LabelEncoder().fit_transform(data['Cuisine'])

默认参数测试集测试准确率:
0.9264281592613964
{'C': 100.0, 'intercept_scaling': 10.0, 'max_iter': 200, 'tol': 0.0001}
0.9604546022763712
调参后测试集测试准确率:
0.9396495781959766

```

调参后的测试准确率可从 0.9264 上升到 0.9396，此处应有掌声~~

4、如果此时找到一家新餐厅，其初始特征情况为评论数 100，人均消费 128，口味 8.8 分，环境 9.2 分，服务 9.0 分，来预测一下这是不是一家好餐厅。

```

best_bread = LogisticRegression(C=100, intercept_scaling=10, max_iter=200, tol=0.0001)
#print(best_bread)
label = pre.Binarizer(threshold=39).transform(bread_data[['Star']])
#print(label)
feature = bread_data[['Comments', 'Per_Consumption', 'Taste', 'Environment', 'Service']].values
#print(feature)
best_bread.fit(feature, label)
is_good = best_bread.predict([[100, 128, 8.8, 9.2, 9.0]])
print(is_good)

```

```

data['Cuisine'] = pre.LabelEncoder().fit_transform(data['Cuisine'])
y = column_or_1d(y, dtype=int)
[1]

```

看来这是一家值得去的餐厅哈~~

那如果有一家[1002, 58, 6.8, 7.2, 7.4]的餐厅呢？

```

is_good = best_bread.predict([[1002, 58, 6.8, 7.2, 7.4]])
print(is_good)

```

```

data['Cuisine'] = pre.LabelEncoder().fit_transform(data['Cuisine'])
y = column_or_1d(y, dtype=int)
[0]

```

矮油~这家餐厅不太推荐哈。。。

四、结论

- 1、上海菜系众多，饮食文化丰富，这也是国际化城市较为重要的一个特征。上海作为原来老法租界，且地理位置靠近浙江和江苏，客户群体较大，所以西餐厅和本帮江浙菜数量分别占据宝座第一、二名也是情理之中。其中最贵的菜系是日本菜，其比较讲究食材的精致和新鲜。最受欢迎的菜系是我国传统美食火锅，反映出国人还是比较喜欢团圆热闹的饮食文化和氛围；
- 2、餐厅星级符合正太分布，3.5 星、4 星和 4.5 星餐厅占了总数量的 90%左右，其中仅 4 星餐厅就占比不到 40%，说明上海整体餐饮水平还是较高的。然后 5 星餐厅数量较少，一方面是餐厅上升空间有限，另外是能够消费得起的群体数量可能不多，毕竟当餐厅水准到一定档次之后，再往上提升的成本、消费也会很高；
- 3、上海餐饮人均消费 108 元，绝大多数消费处于 150 元以下，这是大多数普通阶层人民消费得起的范围，也有相当范围的餐厅消费水平位于千元以上，对标高端消费群体，也反映了社会普遍存在的贫富差距特征；
- 4、整体的市场水平发展还是很健康很平衡的，各项评分、星级分布还有综合评分情况都很均衡，没有明显的劣势指标，这是正常市场化发展的特征；
- 5、利用数据挖掘算法对餐厅做好坏二分类是可行的，在笔者所试验的四种分类算法中，LR 分类方法的测试准确率和计算时间这两项指标最优，选取 LR 算法进行参数调优可将测试准确率提升 1.32 个百分点，同时对未知等级情况餐厅，可利用其特征数值对该餐厅进行预测，判断餐厅是否值得一去。