在线食品交付偏好的探索与预测

李逸萌[1]，管邦祥[2]，吴　羿[1]

1. 南京大学　数学系；2. 上海财经大学　大数据学院

摘要：近几年，伴随着互联网的发展，在线食品配送业务成为了新潮流。在此背景下，本文将对”在线食品交付偏好-班加罗尔地区”数据集开展研究，通过描述性统计建立印度在线食品配送平台消费者的用户画像，利用逻辑回归、决策树和Smote 采样等方式建立模型，研究影响顾客购买意愿的因素，并给出相应的预测。最后，我们将归纳总结上述的观点，并就印度外卖行业的企业 Zomato 的发展战略提出建议。

关键词：在线食品；外卖；逻辑回归

1　背景介绍与研究问题

伴随着技术的革新，印度的在线食品配送行业在过去几年得到了蓬勃的发展，行业中两大头部企业 Swiggy 和 Zomato，在过去的 2020 财年分别有 295.6 和 248.6 亿印度卢比，此外，美国的商业巨头亚马逊也加入了印度外卖平台的竞争。如何占据更大的市场份额是每家企业最为关心的话题。

对消费者的正确认知是市场战略中重要的一环，这能帮助企业了解用户的需求与偏好，调整企业的发展战略，更好地迎合用户，从而在市场竞争中占领有利地位。本文选取了印度班加罗尔地区在线食品配送用户偏好的数据集，旨在研究用户画像以及顾客是否会再次线上下单的影响因素。本文将会通过统计性描述和建立模型分析上述问题，并结合外卖平台 Zomato 的境况提出企业的发展建议。

2　数据说明

本文数据来源于: Data Fountain 网站上的”在线食品交付偏好-班加罗尔地区”数据集。

数据集中包含一个完整的调研问卷和相应的用户反馈结果。问卷共计有效填写量 338 条，包含 55 项与订购用户相关的调研内容。问卷的发放方式为简单随机抽样（simple random），即随机选取到店和在线下单的用户并邀请填写问卷内容。此外，问卷中的经度和纬度记录了受访用户进行在线食品订购的商户位置。

为简化后续程序的运行，我们在数据预处理阶段将部分描述程度的变量转化为了整数，转化方式如下：

表 1　描述性变量转换

变量内容	转换整数
Strongly agree (Very important)	2
Agree (Important)	1
Neutral (Moderately important)	0
Disagree (Slightly important)	-1
Strongly disagree (Unimportant)	-2

数据集的（部分）变量如下：

表 2　描述性变量含义

变量	含义
Output	用户是否愿意再次线上下单
Age	用户的年龄
Monthly income	用户月薪的等级
Ease and convenience	线上下单的便捷程度
Late delivery	较慢的配送对不再购买的影响程度
Politenessy	送餐骑手的礼貌程度
Temperature	食品温度的重要程度

其中 Output 变量将作为本数据集主要的因变量。

3　描述性统计

我们将从用户画像和问卷调查两大部分进行描述性统计。在用户画像中，我们将分析年龄、职业、兴趣偏好、地理信息等维度，提取用户的特征，形成用户标签。在问卷调查中，我们利用灰色关联度分析以及变量的协方差矩阵研究问卷中各指标之间的关联程度。

3.1 用户画像

3.1.1 人口属性

经过对调研问卷中的年龄字段进行统计，我们发现用户的年龄主要集中在 18 岁到 33 岁之间，即用户群体以青年为主，而所有被随机选中参与调研的用户的平均年龄为 25 岁。

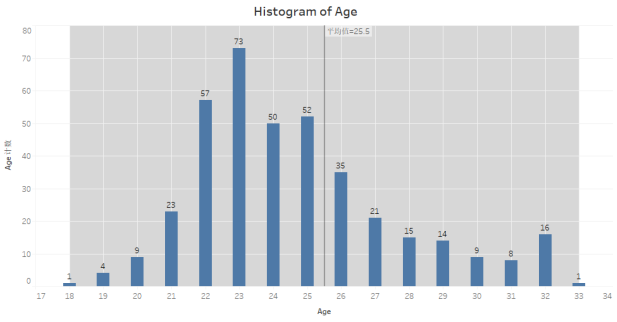

图 1 用户年龄分布

此外，受调研的用户中男性客户的数量多于女性，占总体用户群样本数量的 57.22%（女性用户占比为 42.78%）。

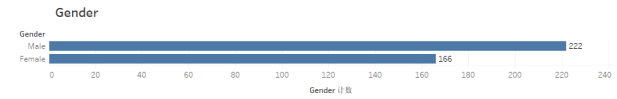

图 2 用户年龄分布

通过分析用户样本中的职业以及月收入构成，我们发现有超过 53% 的用户为学生群体，这与先前分析中的用户年龄集中区间相印证。同时，学生群体的相对占优也部分解释了有超过 50% 的用户月收入小于一万卢比。

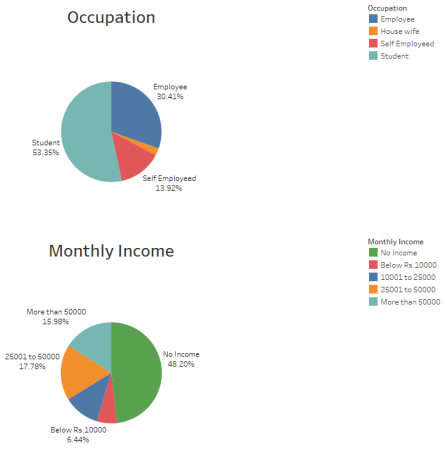

图 3 用户职业和收入分布

3.1.2 兴趣偏好

用户依据自己的 Meal Preference 与 Medium Preference 分别选择了两种偏好，我们将用户的第一选择和第二选择以 0.6 和 0.4 的比重进行打分，得到了如下的结果。用户最喜欢在晚餐时下单, 其次是小食和午餐；超过半数的用户倾向于通过 APP 下单，其次是直接打电话。

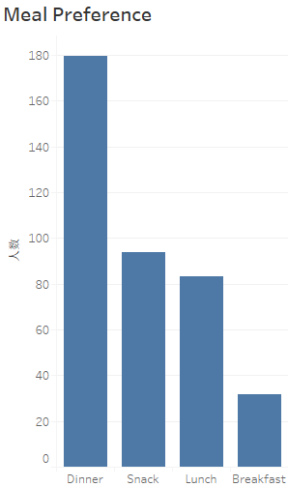

图 4 Meal Preference

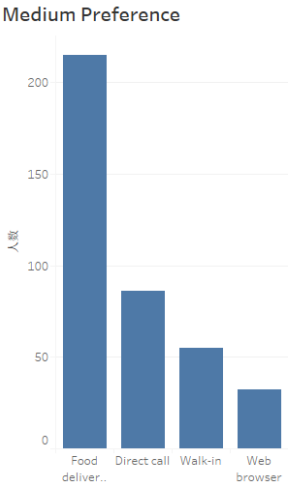

图 5 Medium Preference

3.1.3 地理信息

依据调研问卷中各个订单的配货商户的经纬度信息，我们得到如下图的 GIS 分布:

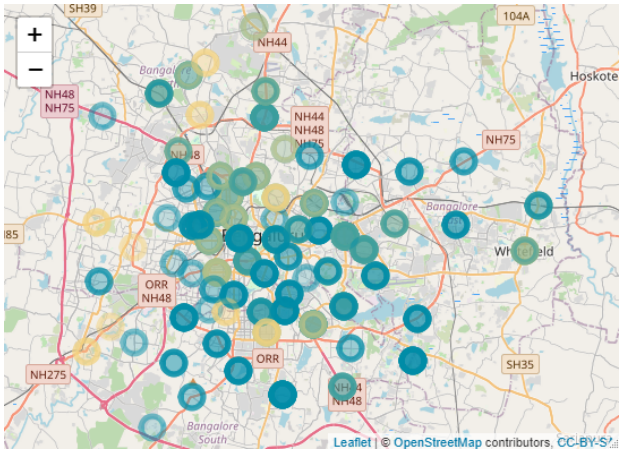

图 6 商户 GIS 分布

详细的 GIS 信息可以在下面的链接中找到:link address 我们可以发现，数据的地理分布与 Bangalore 市的人口密度分布较为匹配，，因此我们认为数据的分布能够较为完整的代表 Bangalore 是的在线事物支付意愿情况。

3.2 问卷调查

3.2.1 问卷信度效度分析

本次使用的数据集主要来源于问卷结果的汇总，因而在对相关项进行具体分析前，我们要对数据的可靠程度进行信度效度分析，结果如下：

Cronbach信度分析-简化格式

项数	样本量	Cronbach α系数
38	50	0.845

图 7 信度分析

效度分析结果 三 排序

名称	因子载荷系数										共同度（公因子方差）
	因子1	因子2	因子3	因子4	因子5	因子6	因子7	因子8	因子9	因子10	
方差解释率%(旋转后)	13.307%	8.232%	7.721%	7.168%	6.695%	6.319%	6.112%	4.657%	4.017%	3.463%	-
累积方差解释率%(旋转后)	13.307%	21.539%	29.261%	36.429%	43.124%	49.443%	55.555%	60.212%	64.228%	67.691%	-
KMO值	0.824										-
巴特球形值	7202.288										-
df	703										-
p 值	0.000										-

图 8 效度分析

在信度分析中，Cronbach α 系数为 0.845，大于 0.8，认为回答具有可靠性；而在效度分析中，KMO 值为 0.824，大于 0.8，表明变量间有较强的相关性，认为回答具有有效性。

3.3 灰色关联分析

灰色关联分析方法是根据被调因素（估值水准）与其余多个因素（指标）之间发展趋势的相似或相异程度，即“灰色关联度”，作为衡量因素间关联程度的一种方法。相应指标 i 的关联系数 r_i 越高表明该指标与估值水准之间的相关性越好。这里，我们将餐厅评级（Output）作为估值水准，选取调研问卷中的部分调研内容为指标进行灰色关联分析。注意到所有指标和估值水准都已是有序因子。

为了消除不同指标原始含义间的区别，我们对所有指标的原始数据进行均质化的无量纲化处理，即所有指标数据除以该指标的样本平均值。经过标准化处理后，我们对两类倾向组分别采用如下公式计算对应倾向中对应指标的灰色关联度 r_i，其中 $\rho \in [0,1]$ 是分辨系数，x_i 是指标 i 中的第 k 个样本对应值，y 对应 Output：

$$\epsilon_i(k) = \frac{\min_i \min_k |y(k) - x_i(k)| + \rho \max_i \max_k |y(k) - x_i(k)|}{|y(k) - x_i(k)| + \rho \max_i \max_k |y(k) - x_i(k)|}$$

$$r_i = \frac{1}{n} \sum_{i=1}^{n} \epsilon_i(k)$$

我们分别选取 $\rho = 0.1$ 和 $\rho = 0.5$ 对所有指标变量进行关联性分析，再选取 $\rho = 0.5$ 对两类倾向中的变量进行关联性分析并得到灰色关联度如下图所示：

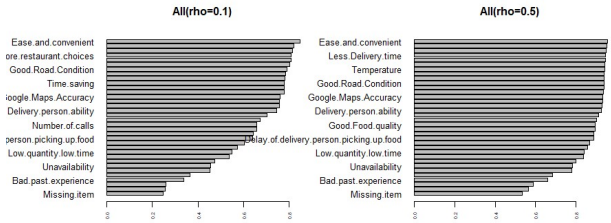

图 9 灰色关联值

图中第一列是选取所有指标变量进行关联性分析后得到的结果。尽管两次计算选取的分辨系数不同，但分析结果中的各指标灰色关联度是高度一致的。我们发现 Ease.and.convenient、Good.Taste 和 Less.Delivery.time 是与用户对商户的评价（Output）高度相关的，而相对的，Missing.item 、Self.Cooking 以及 Wrong.order.delivered 指标则与用户评价的关联度较低。除此以外，其余大部分指标与用户评价反馈的关联性较为相近。

3.4 变量相关性分析

通过上一部分灰色关联度的分析，我们选取了其中其中与 Output 关联度较大的变量 Ease.and.convenient, Good.Taste, Time.saving …. 以及 Age 共 10 个变量，并以 Output 做区分，研究变量之间的相关度及其与 Output 之间的关系，结果如下：

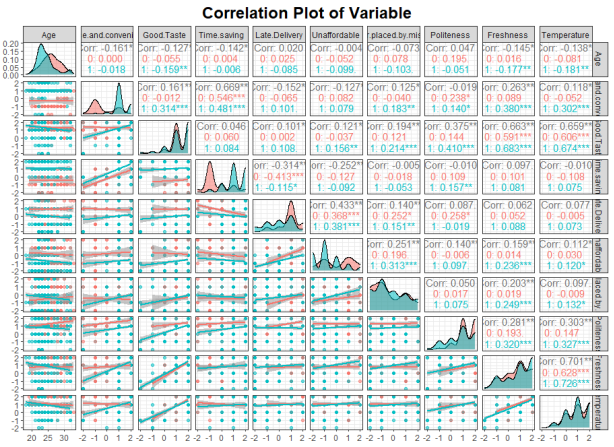

图 10 相关性矩阵

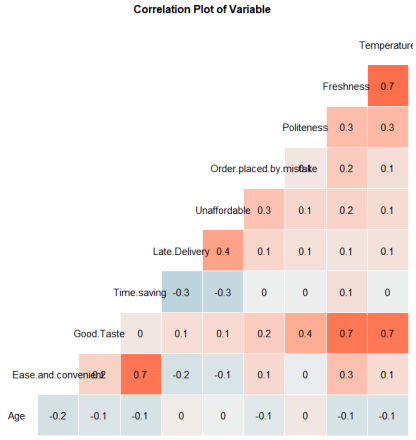

图 11 相关性矩阵

我们发现：

- 年龄与再次购买意愿呈现显著的负相关，说明年龄越大的消费者再次购买的意愿越少；
- 越认为 Ease.and.convenient 很重要的人，越有可能再次购买；
- 越认为 Time.saving 很重要的人，越有可能再次购买；
- 年龄越大的人往往对餐厅个指标的重要性打分越低；
- Time.saving 与 Ease.and.conveniet 有高度相关性；
- Freshness 与 Temperature 有高度相关性；

4 模型的建立与比较

4.1 基于 AIC 的模型初步筛选

由于数据变量较多，为了对预测 Output 的模型有跟好的解释性, 我们首先根据先验知识排除一些变量，利用逻辑回归模型和 AIC 准则初步筛选出显著的九个自变量。

MODEL FIT: $\chi^2(23) = 277.84, p = 0.00$ Pseudo-R^2 (Cragg-Uhler) = 0.78

Pseudo-R^2 (McFadden) = 0.67

AIC = 183.15, BIC = 278.22

表 3 选取显著的 9 个变量以及 Output 因变量

变量	含义
Output	用户是否愿意再次线上下单
Age	用户的年龄
Ease and convenience	线上下单的便捷程度
Time saving	是否节省时间
Late delivery	较慢的配送对不再购买的影响程度
Unaffordable	是否可以负担
Order placed by mistake	食品配送错误
Politenessy	送餐骑手的礼貌程度
Freshness	食品新鲜程度
Temperature	食品温度的重要程度

4.2 训练集测试集的划分

我们随机选取 70% 的数据作为训练集,30% 的数据最为测试集；对一些需要迭代求最优参数的模型 (如 C5.0)，将测试集再划分为训练集 (train_iteration) 和验证集 (val)。

4.3 多模型比较

本报告基于外样本的模型精度评估试验, 对多模型进行训练与预测, 各重复 100 次得到如下表:

模型	ACCURACY TEST	SENSITIVITY TEST	SPECIFICITY TEST	AUC_TEST	AUC_TRAIN
C5.0决策树	92.31%	0.9259	0.922	0.924	
SMOTE LR	70.94%	0.9630	0.633	0.951	
SMOTE LR AIC	80.34%	0.9630	0.7556	0.949	
LR	88.03%	0.6296	0.9556	0.944	
SMOTE AGE^2 LR	83.76%	0.9630	0.800	0.933	
SMOTE AGE^2 AIC LR	88.03%	0.9630	0.8556	0.938	
SMOTE NAÏVE BAYES	88.89%	0.9259	0.8778	0.939	
NAÏVE BAYES	91.45%	0.8519	0.9333	0.942	

图 12 模型综合比较

综上，我们可以发现，就测试集合的准确率而言，C5.0 决策树的准确率最高，为 92.31%，其次是基于 smote 采样的朴素贝叶斯模型, 准确率为 88.89%；

然而，由于他们不是线性模型，这两者模型不具有可观的解释性；

特别的，通过之前的描述性统计与常识可以知道，年龄分布是类似正态的，因此有必要考虑年龄的二次项使得模型更加有效，而上表也进一步证实了引入年龄二次项使得模型的精度有较为显著的提高; 由于篇幅原因，我们接下来将只选取线性模

型中的 smote 采样后的带有年龄二次项模型进行具体解释。

值得注意的是，在具体选择模型的时候，需要结合具体所研究的数据特征进行选择模型；例如，没有 smote 采样的模型由于训练的样本不平衡，在测试集上的准确率较高但是在训练集上的 AUC 较少，容易发生误判，对于未知 Output 的分布未知的数据，则 Smote 的采样显得必要了。

4.4 Smote 采样

我们发现，在原始数据中，顾客是否再一次购买食品的意愿结果 (Output) 是不平衡的，愿意的比例 ($Output = 1$) 是不愿意的 ($Output = 0$) 的 3.46 倍，因此为了使得模型更加完善, 我们对训练数据进行 smote 采样。

4.4.1 寻找 Smote 最优参数 k

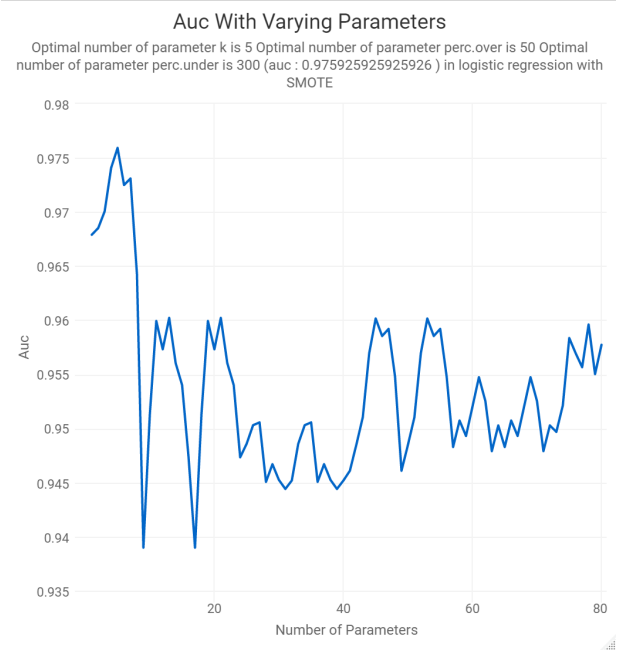

图 13 模型 AUC 随 k 的变化

我们比较多 Smote 参数，选取参数 k=5 使得 Smote 采样的模型在训练集上的 AUC 最大 ($AUC = 0.976$)。

采样后的训练集共有 180 条数据，其中再次购买意愿 ($Output = 0, 1$) 的数据达到平衡，各 90 条。

4.4.2 基于 smote 采样含年龄二次项的逻辑回归模型

由于上述讨论的变量设计用户和人的只有年龄, 如果讨论交互项意义不大, 不妨引入年龄的二次项 Age^2；

	Est.	S.E.	z val.	p
Intercept	67.62	20.03	3.28	0.00
Age	-4.74	1.51	-3.14	0.00
I(Age^2)	0.08	0.03	2.95	0.00
Ease.and.convenient	1.71	0.36	4.71	0.00
Time.saving	0.45	0.28	1.63	0.10
Late.Delivery	-0.75	0.32	-2.33	0.02
Unaffordable	-0.65	0.26	-2.49	0.01
Order.placed.by.mistake	0.53	0.25	2.18	0.03
Politeness	-0.89	0.33	-2.71	0.01
Freshness	0.05	0.30	0.16	0.87
Temperature	-0.40	0.36	-1.12	0.26

模型的解释如下：

我们考虑了年龄的二次项 $I(Age^2)$ 在此模型中共有十个自变量（不含截距项），其中除了新鲜度和温度不显著，省时略微显著，其余所有变量都是十分显著的，具体而言：

1. 年龄越大越不可能再此购买在线食品，相比于没有引入年龄二次项，其回归系数绝对值明显减小了很多；
2. 认为方便度重要的顾客更容易再次购买在线食品，且再次购买的优势比是 $e^1.7$ 倍；
3. 顾客每认为对 Time.saving 的重要性上升一个档次，顾客越可能购买，且顾客购买的优势是不够买的优势的 $e^0.45$ 倍；
4. 对于 Order.placed.by.mistake，回归系数显著为正，说明越认为其重要的顾客越可能再次购买，且重要性每上升一个档次, 再次购买的优势是不购买的优势的 $e^0.53$ 倍；
5. 对于 Late.Delivery，Unafforadable，Politeness 他们的系数显著为负值，说明认为这三个变量重要的顾客，越不可能再次购买；

我们通过观察上一个模型发现其还包含一些不显著的因素，例如 FFreshness 新鲜度和 Tempera-

ture 温度，因此 kao'lü 考虑 AIC 准则选择更加简洁的模型；

	Est.	S.E.	z val.	p
Intercept	66.09	19.75	3.35	0.00
Age	-4.67	1.49	-3.12	0.00
I(Age^2)	0.08	0.03	2.94	0.00
Ease.and.convenient	1.66	0.36	4.66	0.00
Time.saving	0.49	0.28	1.79	0.07
Late.Delivery	-0.66	0.30	-2.18	0.03
Unaffordable	-0.70	0.26	-2.68	0.01
Order.placed.by.mistake	0.49	0.24	2.10	0.04
Politeness	-0.99	0.31	-3.17	0.00

从下图中可以看到，在全模型下 (AIC 选择之前) 与选模型下的 ROC 曲线非常相似，两者 AUC 分别是 96.2% 与 96.6%，

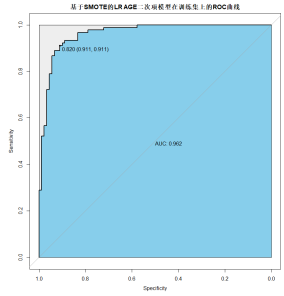

图 14 全模型 ROC 曲线

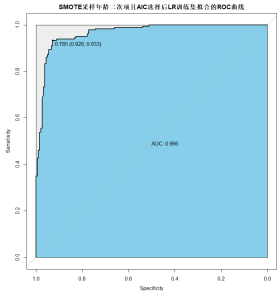

图 15 选择模型 ROC 曲线

因此，们在这里选择 AIC 模型。

5 结论与建议

从描述性统计来看：

最多的群体为单身的男性用户，他们主要由学生和办公人员构成，他们中多数人不分周中与周末都愿意在线下单，他们偏爱在午餐时下单。这样的群体通常具有消费水平较低、追求时间效益的特征。针对这类用户，商家可以选择性价比高、配货速度快的商品向他们推荐。

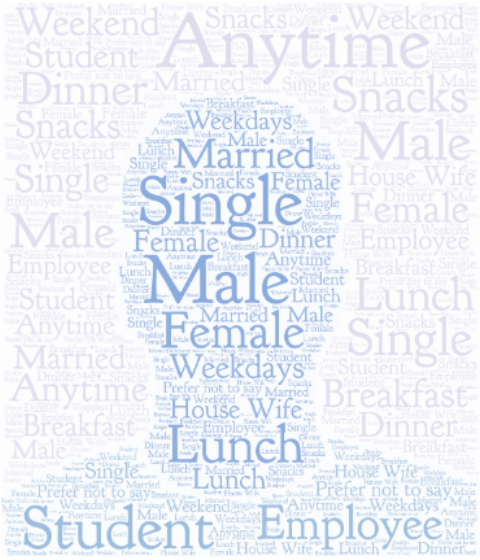

图 16 词云

从模型的角度来看:

对于商家，如果不考虑消费者的背景信息，从问卷结果部分来看，商家的提供食物的便利性十分重要，因此商家应该尽力使得其在线食品配送尽量简化，方便化；

省时这一指标也十分重要，这包括了商家食品制作的速度，配送的速度与配送距离，路线的合理安排与优化；商家可以注重与配送方的及时的有机沟通协调；

另外，错误的食品放置也是较为重要的指标，商家如果想要提高消费者的再次购买率，应该着重避免食品的错误放置；

最后，我们将分析一下 Zomato 是印度在线食品配送平台的第二大企业，结合上述统计分析以及市场环境，我们利用 SWOT 模型进行下，并提出建议:

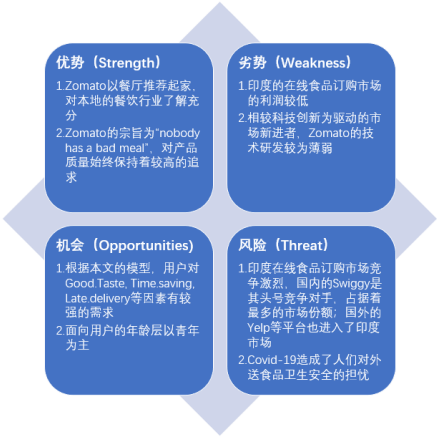

图 17 SWOT 分析

图 18 SWOT 分析

参考文献

[1] 徐轻. 食品网络商店形象对消费者购买意愿的影响研究[D]. 南京农业大学,2017.

[2] Raman. Zomato: a shining armour in the foodtech sector[J]. Journal of Information Technology Case and Application Research, 2018, 20(3-4) : 130-150.

[3] Chandrasekhar Natarajan and Gupta Saloni and Nanda Namrata. Food Delivery Services and Customer Preference: A Comparative Analysis[J]. Journal of Foodservice Business Research, 2019, 22(4) : 375-386.