基于大众点评的上海地区补习班研究

第四组

林耿 王众 张雨晴 陆静颐

一、背景介绍

如今，教育培训市场火爆，从幼儿到学生到成人都是需求个体，已然组成了一个庞大的教育培训市场。“教育培训业是 21 世纪最朝阳产业之一”的观点，已成为越来越多人的共识。

据数据显示，2017 年的教育培训机构数量合计在 10.33 万所左右，总量逐年平稳下降。虽说教育培训机构数量有所下降，但教育培训市场规模呈上升趋势。

图 1 我国教育培训行业市场规模及课外辅导市场规模及预测情况

总体来说，教育培训机构发展潜力巨大，培训产业利润逐年逐步增长，企业规模不断扩大，培训业正以每年 30%以上的速度迅速增长，行业整体发展速度快。据数据显示，K12 培训市场规模从 2012 年的 2281 亿元增至 2017 年的 3930 亿元，年均复合增长率为 9.49%。

随着教育培训机构的发展壮大，所针对的培训内容也逐渐丰富，不仅仅针对于学生升学方面的课业补习，例如音乐、美术等兴趣爱好类补习班也如雨后春笋般高速增长。除此之外，还衍生出针对大学生以及成人的职业规划类培训班。同时，随着外资企业不断进入中国市场，以及留学人数的不断攀升，众多培训机构都纷纷推出语言类课程，并针对不同年龄群体。

我们从大众点评网站爬取了 3000 余条数据，涉及升学、美术、音乐、兴趣、职业、语言六大方面。我们从中删选出 750 条针对语言类培训的有效数据集，试图研究语言类培训班的简介、特色、师资、位置等因素哪些会影响人们在预订补习班时的选择以及体验补习班后的满意度。

二、数据预处理

我们从大众点评网页上爬取得到了 3000 余条数据，并重点处理了语言类补习班的 750 余条数据。数据中的文本数据较多，我们对长句通过分词并提取关键词的形式来进行处理。

图 2 大众点评网页版补课班板块截图

数据清洗过程：

1. 在对教师学历的处理中，我们将其处理为具有 8 水平的哑变量，包含本科、硕士、国际、证书、华东师范大学等；另外，对所有教师的教龄取得均值作为培训班整体教龄的反映。
2. 在对教授科目的处理中，我们通过计算词频，获得各补习班教授最多的三大科目分别是英语、口语、雅思，并将这三个科目作为本次研究的主要对象。
3. 在对特色的处理中，提取高频关键词，最终选取了免费试听、水平测试、一对一等六个水平处理为哑变量。
4. 在对评论的处理中，我们分析了各培训班最具人气的十条评论提取关键词，以各关键词在评论中出现的比例作为各关键词列的值。
5. 由于爬得的项目数据中不直接包含“覆盖人群”这一指标，而这一指标却对培训班的操作有重要影响，通过进一步分析发现，商户介绍中包含这一信息，所以我们对各培训班“商户介绍”的文本提取了关键词，另起“覆盖年龄段”项，分出幼儿、中小学、成人三个哑变量。
6. 评价中的用户反馈包含了效果、师资、环境三项指标，我们对其进行了取均值的操作，以该项均值作为用户满意度的表现。

三、描述分析

1. 变量说明

根据我们爬虫得来的数据，我们将其分类整理如下：我们把预约人数和满意度作为因变量，在接下来的分析过程中，对预约人数做回归分析，而将满意度按照是否超过 8 分做“高”与“低”的分类处理。

变量类型	变量名	详细说明	取值范围
因变量	预约人数	连续变量	[10,9098]
	满意度	连续变量	[6.28,9.4]
自变量	评论数	连续变量	[3,3001]
	分店数	连续变量	[1,43]
	创始年份	连续变量	[1912,2019]
	区域	定性变量(15水平)	静安，黄浦，徐汇，闵行等
	教师学历	定性变量(8水平)	本科，硕士，国际，证书，华东师范大学等
	教龄	连续变量	[0,55]
	教授科目	定性变量(3水平)	英语 / 口语 / 雅思
	特色	定性变量(6水平)	免费试听 / 小班 / 水平测试 / 周学习反馈 / 月学习反馈 / 一对一
	评论	定性变量(9水平)	老师超赞，环境优雅，教学先进等
	覆盖年龄段	定性变量(3水平)	幼儿 / 中小学 / 成人
	价格	连续变量	[19,66486]

图 3 变量说明表

2. 不同种类补习班的统计数据

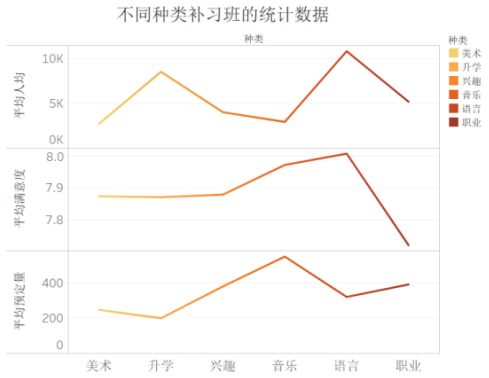

图 4 不同种类补习班统计

对于不同种类的补习班来说，从人均来看，语言类补习班的人均消费要比其他种类高得多，超过了 1 万元。而相比之下美术和音乐类，也就是偏艺术的类别，反而要更便宜些。而从满意度来看，语言类补习班的满意度最高，平均约为 8.0，职业类补习班的满意度较低，大约为 7.8。从预定量来看，音乐类补习班的平均预定量是最高的，而升学类的预定量较低。

3. 创始年份

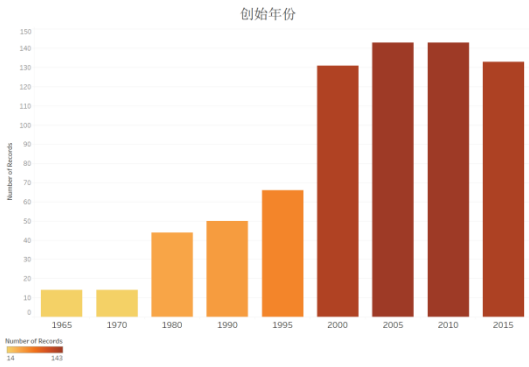

图 5 各语言类补习班创始年份

超过 60%的语言类补习班都是在 2000 年后创立的，而最早有创立于 60 年代的语言类补

习班，大约有 1/20 的补习班创立于六七十年代。在 2000 年后创立的补习班中，各个年份对应的补习班数量都较为平均。

通过查阅资料，我们发现，创立于六十年代的这些语言类补习班大都来自于海外，如早在 60 年代就创建的英孚教育（1965 年 瑞典），而国内的兴起在八九十年代左右，如新东方（1993 年），昂立外语（1984 年）。

4. 分店数

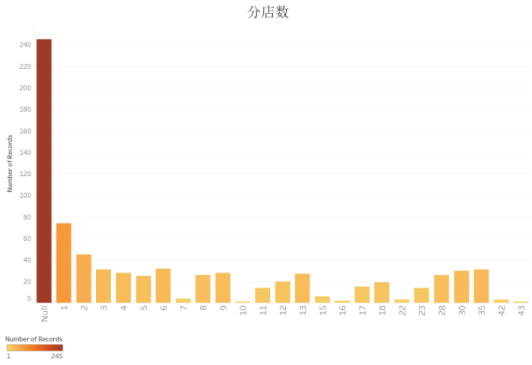

图 6 各语言类补习班拥有分店数分布图

大约 30%的语言类补习班都没有分店；而在有分店的补习班中，最高的拥有 43 家分店，其余分店数对应的补习班数量都比较接近。由于可能会出现一些培训机构拥有多家分店且都出现在统计结果中，因此分店数大于一家的培训机构实际数量会相对更少。

5. 地域分布

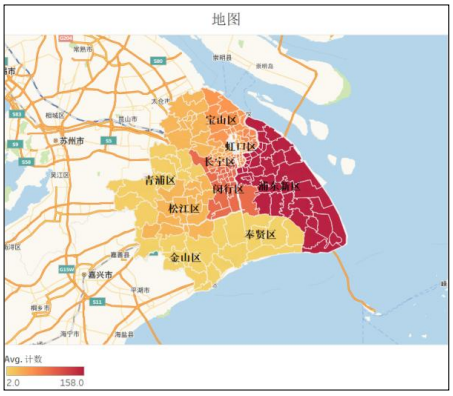

图 7 语言类补习班在上海的区域分布

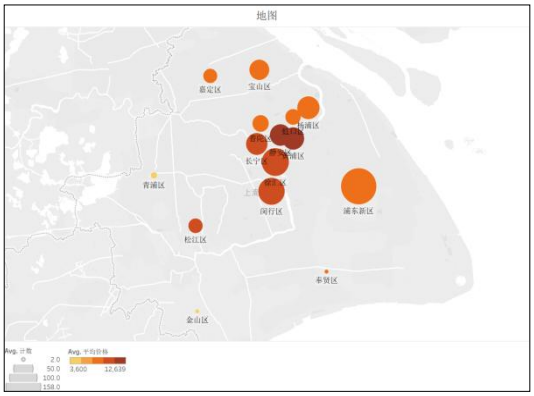

图 8 语言类补习班在上海各区的均价情况

从图中可以看到，位于浦东新区的补习班数量最多（很可能与其面积较大有关），其次是闵行区、长宁区等，位于奉贤区、青浦区、金山区等相对偏远的地区的补习班数量相对较少。

而在这些语言类补习班中，位于黄浦区、静安区的补习班均价最高，位于长宁区、徐汇区的补习班均价也相对较高，而拥有语言类补习班数量最多的浦东新区均价却相对较低。总

体而言平均价格与地域呈现一个放射趋势，中心地段价格较高，距离市中心越远，补习班的价格越低。

6. 补习年龄段

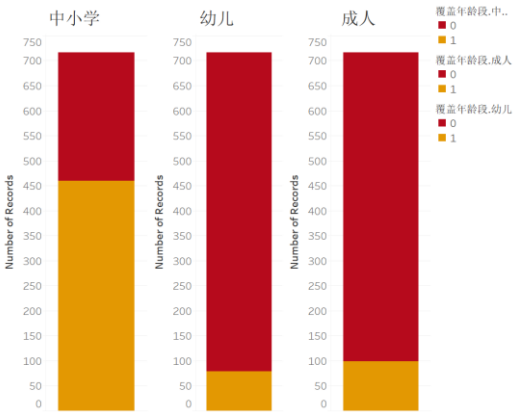

图 9 各语言类补习班覆盖人群情况

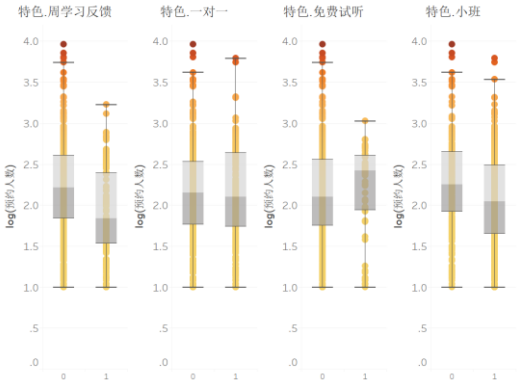

图 10 语言类补习班预约人数与特色的相对关系

从堆叠图（如图 9）中可以看出，中小学覆盖的比率最高，可见，中小学是补习的主力军。

7. 特色

我们提取了一些出现最多的特色，可以看到（如上图 10），有免费试听的预约平均人数大大超过了无免费试听的，这说明，免费试听更吸引消费者；另外，没有周学习反馈的补习班预约人数显著超过有周学习反馈的补习班，因此可以猜测，没有周学习反馈的补习班更受欢迎。

8. 教师学历

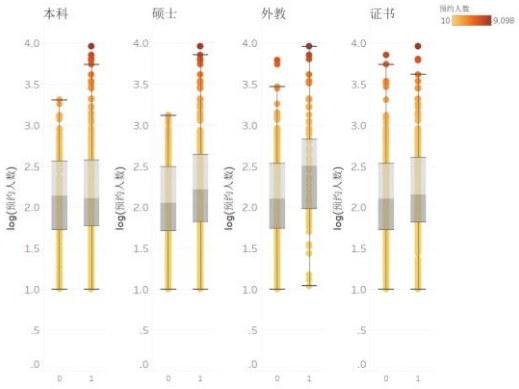

图 11 语言类补习班预约人数与教师学历中关键词的相对关系

我们提取了一些教师学历介绍中出现次数较多的关键词，可以看出，高学历教师和外教对应的预约人数较多，是补习班的招生法宝。

四、模型构建

我们把预约人数和满意度作为因变量，在接下来的模型构建过程中，分别对其做回归分析与分类分析。

对预约人数做回归分析，主要以线性回归为主，通过对变量重要性的分析得到结论；将满意度按照是否超过 8 分做“高”与“低”的分类处理，采用了如逻辑回归、LDA、朴素贝叶斯等等不同的分类模型。

下面选取有代表性的模型做深入分析。

1. 线性回归模型

在线性回归模型中，我们分别使用 log 预定数和满意度作为因变量，评论数、分店数、创始年份、区域、教师学历、教龄、教授科目、特色、评论、覆盖年龄段、价格作为自变量，并在回归分析前都对各自变量的值进行了标准化处理，以便我们更直观的进行系数比较。

- **Log 预约人数线性回归模型分析：**

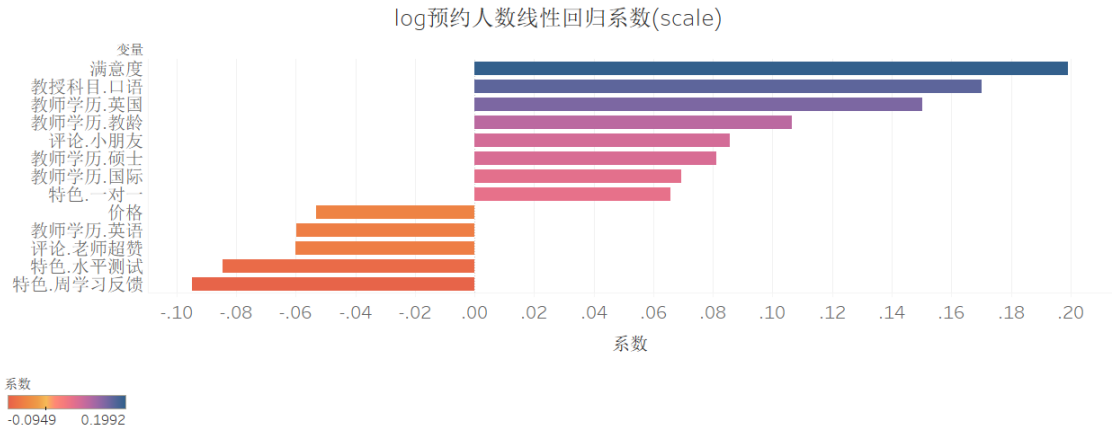

图 12 log 预约人数线性回归结果系数图

① 在经过线性回归变量筛选后，我们可以发现，在与预约人数相关的因素中，满意度、评论数对预约人数有正影响，价格对预约人数有负影响。评论数多的补习班往往关注度和曝光度都很高，因此预约人数也会相对较多；补习班的满意度越高，相应补习产品的好评越多，也越容易被客户之间相互推荐，自然预约人数更多；而人们在选择补习班的时候毫无疑问会考虑性价比的因素，因此价格高的补习班预约人数相对更少。

② 教师学历中的教龄、硕士、国际、英国都对预约人数有正影响。不难理解，教龄长的、以及有着硕士留学方面背景的教师，相对经验较为丰富，也更收到同学家长的青睐，预约人数更多。

③ 从回归结果中我们发现，课程特色包含一对一的课程相比水平测试与周学习反馈的预约人数更多。可见，家长学生更倾向于选择一对一的语言类补习班，确实在学习语言的时候一对一的教授环境更利于学习和使用一门新的语言，因而预约人数较多。

- **满意度线性回归模型分析：**

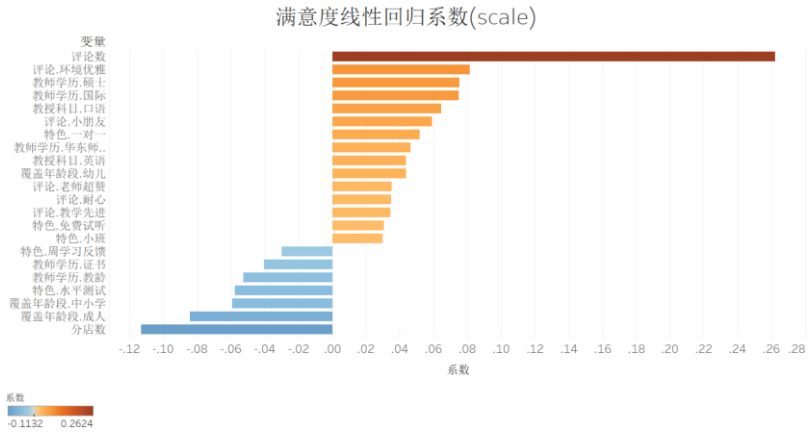

图 13 满意度线性回归结果系数图

① 在经过线性回归变量筛选后，我们可以发现，在与满意度相关的因素中，评论数与各种好评都对满意度有正影响。不难理解，评论多意味着高曝光，好评数多意味着课程的体验效果较好，自然满意度较高。

② 分店数对于满意度则有着负影响。分析原因，可能是因为随着分店的增加，对于各分店的管理难度逐渐升高，师资水平、教学环境等难以维持在同一水准，各分店的参差标准容易引发家长分歧，导致满意度的下降。

③ 针对课程特色，小班、一对一、免费试听对满意度有着正影响，而周学习反馈、水平测试有着负影响。可见，提供免费试听的课程往往水准较高，因而有自信给到试听，满意度也会较高。小班和一对一的教学模式会带来较好的教学成果，满意度更高。而周学习反馈与水平测试相对更死板与墨守成规，影响了满意度。

④ 覆盖年龄段中，幼儿对满意度有正影响，而中小学生与成人则有负影响。这可能是因为幼儿的年纪较小，尚未形成自己的思考模式，更容易给出好评，随着年龄的增长，逐渐有了自己的思考，更容易发现课程中包括教学模式、课程设计、环境设施等方面的不足，造成满意度的下降。

2. 逻辑回归模型

由于不同人针对一样客观事物的评价标准可能不同，而一般而言，8/10（四星及以上）可以作为大部分人群评价事物好坏的指标，因此我们将满意度这个连续变量以 8 分为分界线分成高、低两类后，建立逻辑回归模型进行分析。

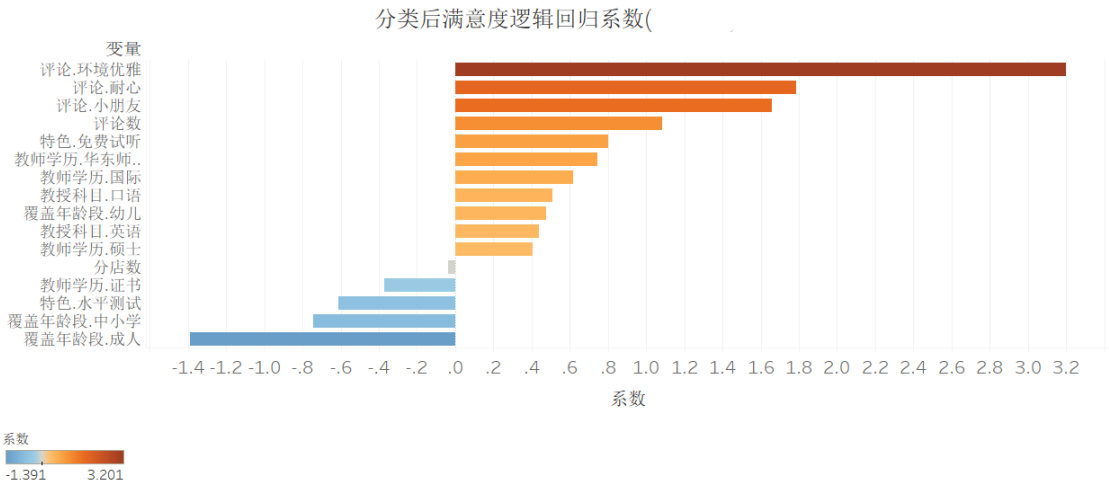

图 14 分类后满意度逻辑回归结果系数图

由图可得：

① 与线性回归模型相同，评论数与好评都会对满意度产生正影响。好评中带有环境优雅、小朋友、耐心的重要性更高。

② 课程特色中，免费试听和水平测试的重要性更高，前者产生正影响，后者产生负影响。

③ 教师学历与覆盖年龄段的模型结果与线性回归基本相同。

为了对包含正影响和负影响在内的各变量的重要性有直观了解，我们又据此做了因子重要性分析，结果如下。

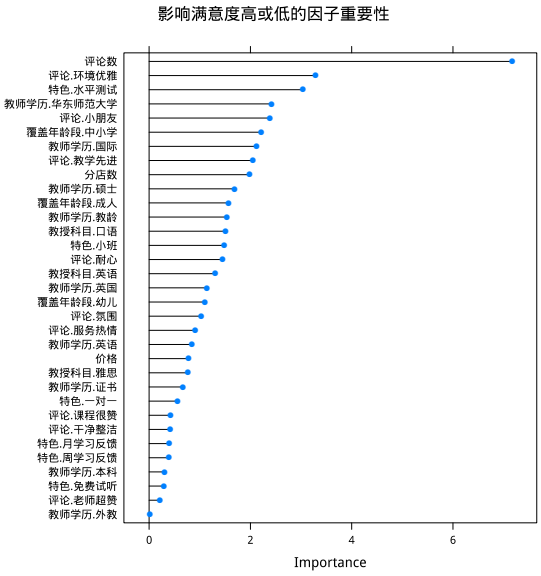

图 15 影响满意度高与低的因子重要性

从图中可以看到，在各因子中，评论数最为重要，可以猜测，消费者在为听课体验打分时，会受到课程是否受欢迎的影响（通过评论数高低表现出来）；在其他因子中，评论中出现环境优雅关键词的比例，特色中是否有水平测试等都有比较高的重要性。

3. 不同分类模型的分类效果

	LDA	Naïve Bayes	Logistic
F1	0.760	0.795	0.710
AUC	0.892	0.915	0.770

图 16 不同分类模型的分类效果

通过不同模型的构建和 AUC 及 F1 值的分析，可以看到在本问题中，朴素贝叶斯分类器的分类效果最好，其 AUC 值甚至可以达到 0.9 以上。

4. 人群关注的差异

已修读人群	未修读人群
分店数	*
教师学历.教龄	教师学历.教龄
教师学历.硕士	教师学历.硕士
教师学历.证书	*
教师学历.国际	教师学历.国际
教师学历.华东师范大学	*
教授科目.英语	教师学历.英语
教授科目.口语	教授科目.口语
特色.免费试听	*
特色.小班	*
特色.水平测试	特色.水平测试
特色.周学习反馈	特色.周学习反馈
特色.一对一	特色.一对一

已修读人群	未修读人群
评论.老师超赞	评论.老师超赞
评论.环境优雅	*
评论.教学先进	*
评论.小朋友	评论.小朋友
评论.耐心	*
覆盖年龄段.幼儿	*
覆盖年龄段.中小学	*
覆盖年龄段.成人	*
*	满意度
*	评论数
*	价格
*	教师学历.英国

图 17 人群关注差异因子表

为了对已修读课程的人群和未修读课程的人群关注因素的差异有所了解，我们做了分析。在取两人群各自因变量的时候，可以分别以满意度和预约人数作为反馈指标，因为这两指标是两人群在接受信息（上课体验或网站呈现的信息等）后作出的直接反应，最后再分别进行自变量的筛选，得到因子表，我们发现，

- **已修读人群对于课程评价更关注：**

是否是小班化教学，环境是否干净整洁，教学是否先进，这些因素都是只有在体验过课程教学才能得到的感受，也是直接影响已修读人群对课程评价的因素，我们也能从中看出，人们在上语言类补习班时会比较注重小班化，环境是否整洁等等因素。

- **未修读人群对于做决策更关注：**

满意度，评论数，价格，教师的海外经历；由于未修读人群了解课程的渠道只是官网，所以

作为能够从官网上直接提取的元素，满意度、评论数、价格、教师的海外经历也成了人们做出“是否预约”这一决策的影响因素，从中也能发现，人们倾向于认为评论数多的课程更有可能是自己会青睐的课程，而教师的海外经历也是影响决策的重要因素，这些因素在学生上课时反而变得不那么重要。

五、结论建议

在经过了上述分析后，我们从商家与消费者两个角度提出建议：

- **商家如何做好补习班？**

我们主要为商家提出四点建议：

口碑带来人气；进军幼儿教育；

教师海外背景；环境决定成败。

① 课程特色中，对于一家语言类补习班来说，口碑好才能带来更好的发展。只通过削减成本而带来收益并不能让补习班得到良性的发展。

② 成年人的生意更难做，幼儿语言补习班往往更容易令顾客满意。不仅如此，带有幼儿教育的语言补习班，事实上也能吸引人预约课程。

③ 对于语言类补习班来说，家长或者成年学生，非常看重教师的国外背景。这些老师在家长和成年学生的心中与优秀的口语教育有着莫大的关系。不管是为了吸引新顾客报名还是提高顾客的满意度，下足本钱引进外教或着有国外受教育背景的教师都是有必要的。

- **消费者怎么挑选补习班？**

消费者在挑选补习班时需要明确三个不等于，即：

高学历教师≠宽心；

资历老教师≠放心；

大品牌连锁≠安心。

① 教师的高学历会吸引更多人参加课程，但是却不一定让人满意，有时甚至会让满意度下降。对于消费者而言，合适的才是最好的，高学历的老师不一定更适合教学。

② 小店家可能更让人满意。从分析来看，分店多的补习班很容易让顾客满意度降低，在挑选补习班时，应该多考虑店少而精的补习班。