论文编号：A0659

西安市少儿编程市场统计测度研究

论文题目:西安市少儿编程市场统计测度研究

参赛学校:西安财经大学

参赛成员(作者):朱天翊、党映天、党洁玉

指导老师:马金萍

摘 要

“互联网+”时代，世界各国都在争先恐后地发展编程教育，中国也出台了很多相关政策和文件强调编程教育的重要性，并且提出要将编程纳入教育体系当中。“少儿编程”也越来越火热，与此同时，越来越多的资本瞄向了这一片新的蓝海。

以西安市的家长对于少儿编程培训机构的需求偏好为出发点，在西安市的不同区县对涉及到的学生群体、家长以及少儿编程培训机构的从业人员进行调查和访问。采取线上和线下两种方式，其中线上为主。根据预调查结果，对于样本收集量较少的地区进行线下问卷发放，同时对相关群体进行质性访谈。调查涉及到消费者对于编程以及在升学方面的相关政策的了解程度、培训机构的课程设置和内容的偏好、少儿编程的线上教学模式的建议等方面的内容，针对调查结果给出建议；对西安市热度较高的三所少儿编程培训机构进行结构式的深度访谈，了解目前西安市少儿编程培训机构的实际情况；利用网络爬虫手段爬取相关点评软件的评论。在定性和定量分析相结合的基础上，结合文本挖掘等方法，采用对学生群体、家长以及少儿编程培训机构的从业人员的调查结果建立随机森林模型、多元logistic回归模型、对比分析和主成分分析的方式进行研究。

调查结果显示，大部分家长已经认识到了少儿编程教育的重要性，同时给孩子报名过其他培训班的家长也更愿意给孩子报名少儿编程培训课程；受访者最关心培训机构的教学质量，同时希望能够增加课堂的师生互动频率。由于编程的特殊性，受访者希望培训机构能够增加较多的实践环节；在对于老师的要求方面，受访者更为看重教师的编程能力而不是迷信“学历”；大多数受访者都更加愿意接受朋友的宣传，而不是大量的广告。

根据以上结果，我们给出了培训机构发展策略的一些建议。第一，培训机构应着眼于提高教学质量，建立交互式平台，增加师生互动；第二，少儿编程培训机构可与其他培训机构进行联合宣传，增强宣传效果；第三，培训机构不应该迷信教师学历，而更应该考察教师的实际编程能力；第四，培训机构应该建立较为完善的教师考核制度，提升机构的实力。希望通过本次调查，可以给予西安市的少儿编程培训机构一些发展思路，为少儿编程教育尽一份绵薄之力。

关键词：少儿编程 随机森林 市场需求 网络爬虫 文本挖掘

Abstrat

In the era of "Internet +", countries around the world are scrambling to develop programming education. China has also issued a lot of relevant policies and documents to emphasize the importance of programming education, and proposed to include programming in the education system. "Programming for Kids" is becoming more and more popular, and at the same time, more and more capital is aiming at this new blue ocean.

Taking parents' preference for children's programming training institutions as the starting point, the student groups, parents and practitioners of children's programming training institutions involved were investigated and interviewed in different districts and counties of Xi 'an. Adopt online and offline two ways, online is the main one. According to the results of the preliminary survey, offline questionnaires were issued to areas with small sample collections, and qualitative interviews were conducted with relevant groups. The survey involves consumers' understanding of programming and related policies on entering a higher school, the preferences of training institutions' curriculum Settings and content, and suggestions on online teaching mode of children's programming. Suggestions are given based on the survey results.Three children's programming training institutions with high popularity in Xi 'an were interviewed in depth in a structured way to understand the actual situation of the children's programming training institutions in Xi 'an. Use web crawler to get the comments of related review software. Based on the combination of qualitative and quantitative analysis, combined with text mining and other methods, this paper adopts the survey results of students, parents and employees in children's programming training institutions to establish random forest model, multiple logistic regression model, comparative analysis and principal component analysis.

The survey results show that most parents have realized the importance of programming education for children, and parents who have signed up for other training classes are more willing to sign up for programming training courses for children. The interviewees were most concerned about the teaching quality of the training institutions and hoped to increase the frequency of teacher-student interaction in the classroom. Due to the particularity of programming, interviewees hope that training institutions can add more practice links; In terms of teachers' requirements, interviewes pay more attention to teachers' programming ability rather than academic qualifications. Most interviewees are more willing to accept publicity from friends than a lot of advertising.

According to the above results, we give some suggestions on the development strategy of training institutions. First, training institutions should focus on improving the quality of teaching, establish interactive platforms and increase the interaction between teachers and students. Second, children's programming training institutions can carry out joint propaganda with other training institutions to enhance the

propaganda effect; Third, training institutions should not trust teachers' academic qualifications, but should examine teachers' actual programming ability. Fourth, training institutions should establish a more perfect teacher assessment system to enhance the strength of the institutions. Hope that through this survey, we can give some development ideas to the children's programming training institutions in Xi 'an, and make a modest contribution to the children's programming education.

Keywords:Children's programming Random forest The market demand Web crawler Text mining

目 录

图目录

表目录

一、绪论

（一）选题背景

近年来，随着国家教育政策的支持，以及资金的大量投入，少儿编程教育前景十分光明。

尽管现阶段少儿编程教育发展基础良好，但同时也暴露出了很多问题:家长对少儿编程的看法、家长希望孩子通过学习少儿编程的收获……以上问题均有待求证。

图 1　少儿编程培训机构发展历史[1]

随着丝绸之路经济带的发展，国家又明确把西安定位为国际化大都市无疑更为西安的发展加速。同时在创新驱动发展工程上国家给予大量投资，不仅显示了国家对西安发展的重视，也显示出国家对科技教育的重视。西安市现有编程培训机构13家以上，包含编程猫、小码王、童程童美等。

《2017-2023年中国少儿编程市场分析预测及发展趋势研究报告》[2]估计国内少儿编程市场规模或达百亿。幼儿园、小学和中学生是编程教育的核心授课群体，据教育部统计，2015年这三类群体的在校人数约为 1.83 亿人；据《2019年全国教育事业发展统计公报》[3]统计，2019年这三类群体的在校人数约为2亿人。据

调查统计中国大陆少儿编程渗透率为0.96%，预计每人每年在编程培训领域消费6000元，粗略估计目前国内少儿编程市场规模达105亿元，而且每当渗透率提升1%，市场规模就有望扩大100亿，发展前景广阔，市场规模巨大。

编程作为互联网、人工智能等各种高新技术的基础和核心，并且随着国家对创新驱动战略对科技人才发展的需求在逐年上涨，其地位不言而喻。据《国家统计年献》[4]显示，我国服务性消费与教育文化娱乐支出消费连年上升，教育文化娱乐占消费支出总比也连年上升，随着家长对于教育重视程度的日趋提高，以及家长对于相关政策的深入了解，越来越多的家长将孩子的培养目标集中在提高孩子未来竞争力上。编程作为培养孩子综合能力的学科，其发展也迎来契机。

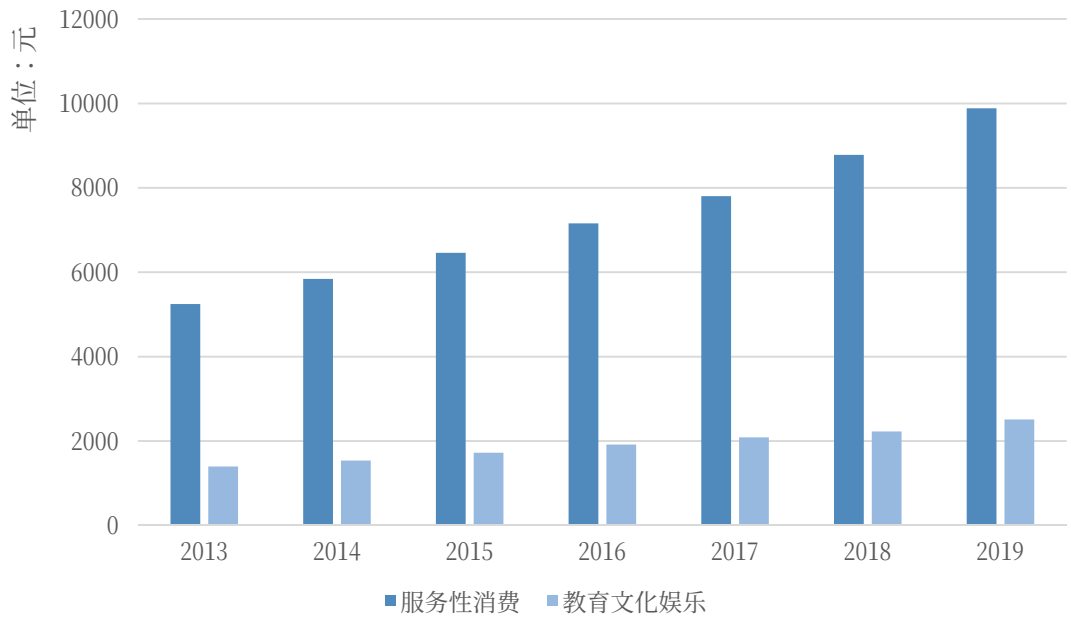

图 2　服务型消费和教育文化娱乐支出

（二）研究路线

图 3　调查的具体技术路线和总体框架

（三）研究目的

1.受访者对少儿编程以及相关政策了解程度

通过问卷调查了解受访者对于少儿编程、编程在未来教育中的发展趋势和编程在升学方面的政策的了解程度，以此来判断西安市家长对于编程这一未来教育的发展趋势的认知情况，探寻市场潜力。

2.受访者对培训机构宣传发展等方面的改进意见

根据受访者对于少儿编程了解方式的相关调查，客观看待目前培训机构所采用的宣传手段，并且结合其他分析结构给出相应建议，帮助培训机构提高宣传效率，扩大消费群体。

3.受访者对培训机构的关注层面和消费偏好

本调查旨在了解西安市的家长群体对于培训机构的课堂设置、教学方式、课程时间安排、教师需求、机构特点的要求以及线上教学的看法，从而更加理性和客观地给予培训机构相应的发展建议。

二、调查设计与组织实施说明

（一）调查目的

本项目采用抽样调查，通过对西安市各区县少儿编程市场发展现状的分析从而推断西安市整体的发展状况，明确整个少儿编程市场的需求并确定培训机构的发展新方向和新目标。抽样调查的目的在于根据样本调查的结果来推断总体，在此次的调查中我们旨在找出少儿编程市场现有的问题，以及消费者在选择培训机构时的评判准则，准确地对少儿编程这一潜力巨大的市场进行深入的调查，并分析出少儿编程市场的需求特征及竞争点所在，力求做到对市场在未来几年的最新动向有一个客观的把控。

（二）调查内容

本次调研的主要内容包括：西安市学生家长对少儿编程及其相关政策的认知程度；西安市学生家长给孩子报名少儿编程课程的主要原因；西安市学生家长对少儿编程培训机构的关注点；西安市学生及其家长对机构授课老师的选择要求；西安市学生家长对培训机构授课方式的主观倾向；西安市学生及其家长对少儿编程培训机构的建议。

（三）调查对象

1.现居住地位于西安市内的学生家长；

2.正在接受编程教育的学生；

3.西安市内发展较为成熟的三所少儿编程培训机构（童程童美少儿编程教育机构、乐博乐博教育、西安小码王少儿编程培训中心）。

（四）调查方式及调查过程

本次调查以问卷调查为调查手段，具体调查过程如下：

采用线上电子问卷发放与线下纸质版问卷发放两种方式进行调查。问卷内容循序渐进，避开了部分可能会触及到受访者隐私的问题并且调查组成员与指导老师均参与了调查问卷的前期设计与审查工作，确保问卷的合理性。由于网络问卷调查人群地区受限，部分地区样本收集量较少，不符合调查要求，因此本调研小组选择在样本收集量较少的地区进行线下纸质版问卷的发放。

（五）抽样设计

1.抽样方法

本次调查主要采用非概率抽样方法。通过预调查问卷回收中地区占比情况，发现高陵区和阎良区回答问卷的样本量远低于其他区县，因此为了更加全面地反映情况，使得预测能够更好地反映西安市的全貌，我们选择在此两个地区额外发放纸质版问卷，用以弥补该地区电子版问卷回收可能不足而导致样本量过少的缺陷，使得我们的调查更加客观和准确。

在样本抽取中发现，部分问卷的回答时间远低于平均时间、主观题内容只是单纯地以数字形式呈现，并没有做出有效的回答。故依据非概率抽样方法剔除了部分不符合样本要求的问卷。

2.样本量的确定与分配

据陕西统计局2020年《陕西统计年献》[5]可得西安2019年少年儿童数量约为200万,根据《问卷统计分析实务——SPSS操作与应用》[6]可得出调查样本数量计算公式：

$$n \geq \frac{N}{\left(\frac{\alpha}{k}\right)^2 \frac{N-1}{P(1-P)}+1} \tag{1}$$

将显著水平α设为0.05，当置信度为1－0.05=0.95时，k=1.96，p=0.5，代入上述公式，计算可得抽样样本数为384。将显著水平α设为0.04，当置信度为0.96时，计算得抽样样本数为600。考虑到样本的准确性，本次调查共发放问卷500份，共回收问卷423份（其中包含电子版问卷366份，纸质版问卷57份）。问卷回收率84.60%，其中有效问卷共346份（电子版302＋纸质版44份），问卷有效率达81.80%。发放问卷用于调查西安市家长对于现如今西安市少儿编程相关内容的认识、期望及意见等相关问题。

三、关于西安市少儿编程市场的问卷调查结果的统计测度研究

（一）Logistic回归模型

1．模型的建立

在研究孩子是否接受过少儿编程培训的问题时，因变量本身只取 0，1 两个离散值，不适合直接作为回归模型中的因变量。针对 0-1 型因变量产生的问题，我们对回归模型进行改进，选用 Logistic 回归模型对调查数据进行分析。

Logistic 函数形式为：

$$f(x) = \frac{e^x}{1+e^x} \tag{2}$$

2．模型的求解

以孩子是否接受过少儿编程培训为因变量（1 为是，2 为否），以家长性别、家长年龄、家庭住址所在地区、家长的受教育程度、家长的月收入、家长的职业、家人是否从事 IT 行业、孩子的受教育程度以及孩子是否在上补习班为自变量进

行二元 logistics 回归拟合。

表 1　模型摘要

-2 对数似然	考克斯-斯奈尔 R 方	内戈尔科 R 方	卡方	显著性
242.529[a]	0.496	0.661	237.025	0.000

由上表可知，卡方值为 267.025，p 值为 0.000，小于 0.05，通过了显著性水平为 5%的显著检验，由此可知模型具有统计学意义。

由模型摘要表可知-2 对数似然值 242.529，考克斯-斯奈尔 R 方为 0.496，但是与线性回归模型不同，在 Logistic 回归模型中考克斯-斯奈尔 R 方、内戈尔科 R 方意义不大，因此不予关注。-2 对数似然值是模型评价的重要指标，本模型中-2 对数似然值为 242.529，说明模型拟合优度较好，除去量化评价拟合优度效果外，我们还对其利用霍斯默-莱梅肖检验进行质性评价拟合优度效果。

表 2　霍斯默-莱梅肖检验

卡方	自由度	显著性
4.816	8	0.777

由霍斯默-莱梅肖检验结果可知，卡方值为 4.816，p 值为 0.777 大于 0.05，说明模型充分利用现有的信息最大化地拟合模型，解释了模型的变异。

3．结果分析

表 3　回归参数表（1）

	B	标准误差	瓦尔德	自由度	显著性	EXP(B)
性别(1)	1.064	0.426	6.233	1	0.013	2.897
家庭住址所在区			18.712	10	0.044	

雁塔区(6)	1.779	0.889	4.003	1	0.045	5.923
临潼区(8)	2.463	0.886	7.721	1	0.005	11.739
孩子是否在上补习班(1)	-2.865	0.524	29.949	1	0.000	0.057
常量	-20.449	40191.221	0.000	1	1.000	0.000

表 4　回归参数表（2）

	EXP(B)	EXP(B) 的 95% 置信区间	
		下限	上限
性别(1)	2.897	1.257	6.676
家庭住址所在区			
雁塔区(6)	5.923	1.037	33.835
临潼区(8)	11.739	2.066	66.700
孩子是否在上补习班(1)	0.057	0.020	0.159
常量	0.000		

由表可得，性别中性别为男（1），家庭住址所在区为雁塔区（6）、临潼区（8）以及孩子在上补习班（1）均对孩子是否接受过编程教育存在显著的影响（$p<0.05$）。其中，性别中性别为男（1），家庭住址所在区为雁塔区（6）、临潼区（8）对孩子接受过编程教育存在正影响，而孩子在上补习班（1）对孩子接受过编程教育存在负影响。

性别中性别为男（1）的系数为 1.064，wald 卡方值为 6.233，p 值为 0.013，小于 0.05，通过了显著水平为 5%的显著性检验。OR 值为 2.897，由此可知在其他变量保持不变的基础上，性别为男的样本，相对于性别为女的样本，在孩子接受过编程教育的概率上要高出 1.897 倍（2.897-1）。

家庭住址所在区为雁塔区（6）的系数为 1.779，wald 卡方值为 4.003，p 值为 0.045，小于 0.05，通过了显著水平为 5%的显著性检验。OR 值为 5.923，由此可知在其他变量保持不变的基础上，家庭住址所在区为雁塔区的样本，相对于家庭住址所在区为其他区的样本，在孩子接受过编程教育的概率上要高出 4.923 倍（5.923-1）。

家庭住址所在区为临潼区（8）的系数为 2.463，wald 卡方值为 7.721，p 值为 0.005，小于 0.05，通过了显著水平为 5%的显著性检验。OR 值为 11.739，由此可知在其他变量保持不变的基础上，家庭住址所在区为临潼区的样本，相对于家庭住址所在区为其他区的样本，在孩子接受过编程教育的概率上要高出 10.739 倍（11.739-1）。

孩子在上补习班(1)的系数为-2.865，wald 卡方值为 29.949，p 值为 0.000，小于 0.05，通过了显著水平为 5%的显著性检验。OR 值为 0.057，由此可知，在其他变量保持不变的基础上，孩子在上补习班的样本相对于孩子未上补习班的样本，孩子接受过编程教育的概率要小 94.3%（0.057-1）。

根据 wald 卡方值可知，孩子在上补习班（1）对孩子是否接受过编程教育的影响最大，其次是家庭住址所在区为临潼区（8），第三是性别中性别为男（1），第四是家庭住址所在区为雁塔区（6）。

根据系数表构建模型如下：

Logit(p)=-20.449+1.064*男+1.779*雁塔区+2.463*临潼区-2.865*孩子在上补习班

4．模型的预测

表 5　分类表[a]

实测	预测

			孩子是否接受过编程教育		
			1	2	正确百分比
步骤 1	孩子是否接受过编程教育	1	158	18	89.8
		2	32	138	81.2
	总体百分比				85.5

以 0.5 的概率为分界线，采用上述模型进行预测，结果如上表所示，其中预测为孩子接受过编程教育，现实接受过编程教育的样本有 158 位，预测正确率为 89.8%；预测为孩子未接受过编程教育，现实未接受过编程教育的样本有 138 位，预测正确率为 81.2%，整体预测正确率为 85.5%。由此可见模型预测结果非常理想。

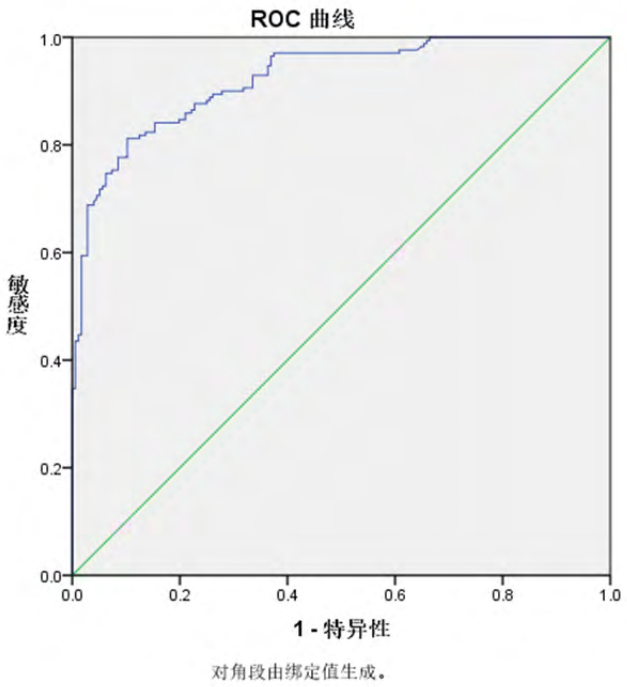

图 4　ROC曲线

表 6　曲线下方的区域

检验结果变量：　预测概率

			渐近 95% 置信区间	
区域	标准误差[a]	渐近显著性[b]	下限	上限
.923	.014	.000	.896	.950

检验结果变量 预测概率 至少有一个在正实际状态组与负实际状态组之间的绑定值。统计可能有偏差。

a. 按非参数假定

b. 原假设：真区域 = 0.5

以预测概率与真实值构建 Roc 曲线如上图所示，其下方所包面积为 0.923，说明该模型对于目标变量具有很好的预测效果。

（二）基于随机森林模型的影响消费者选择的主要因素研究

1.随机森林模型介绍

随机森林通过对决策树做去相关处理，实现对装袋法树的改进。在随机森林中需对自助抽样训练集建立一系列决策树，但是在建立决策树时，每考虑树上的一个分裂点，都要从全部的P个预测变量中选m个变量的随机样本作为候选变量。每个分裂点所考虑的预测变量个数约等于预测变量总数的平方根。

利用自助法从原始数据集中抽取大约为 70%的样本作为模型的训练集，未抽到的 30%的数据作为试验集。通过对抽取的训练集数据建立不同的决策树模型，并采用简单多数投票法对不同分类结果进行投票选出最终的分类结果。

2.变量预处理

① 类别标签的确认。将“您是否给孩子报名过少儿编程培训班”这种二元选择行为表示为分类标签，当选择“是”时，取值为 1;否则，取值为 0。

② 将问卷中年龄的分组，地区等属性特征值，按照问卷中的选项顺序对应的特征属性的具体内容进行相应转化。

表 7 变量名称与含义

变量名称	变量含义
Q1	性别
Q2	年龄
Q3	家庭所在地
Q4	学历
Q5	月收入
Q6	工作类型
Q7	家人是否有从事互联网工作
Q8	孩子所处的教育阶段
Q9	孩子是否在上其他类型的培训班
Q10	是否给孩子报名过少儿编程培训课程

3.随机森林模型的构建

我们根据经验将初始的决策树数量设置为 500，并且计算模型每个变量的特征重要性。并画出最有效的决策树，对于试验集进行分类计算基于基尼系数和准确率的模型正确率。

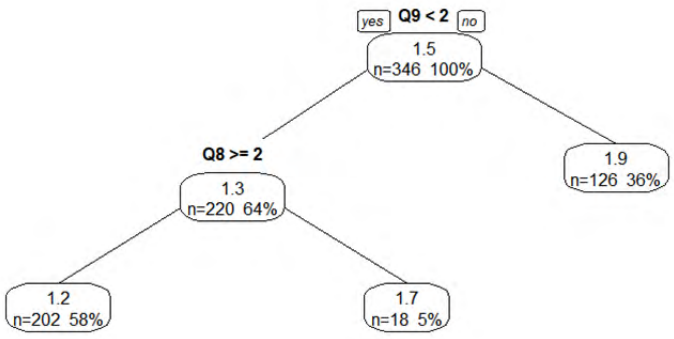

图 5　决策树

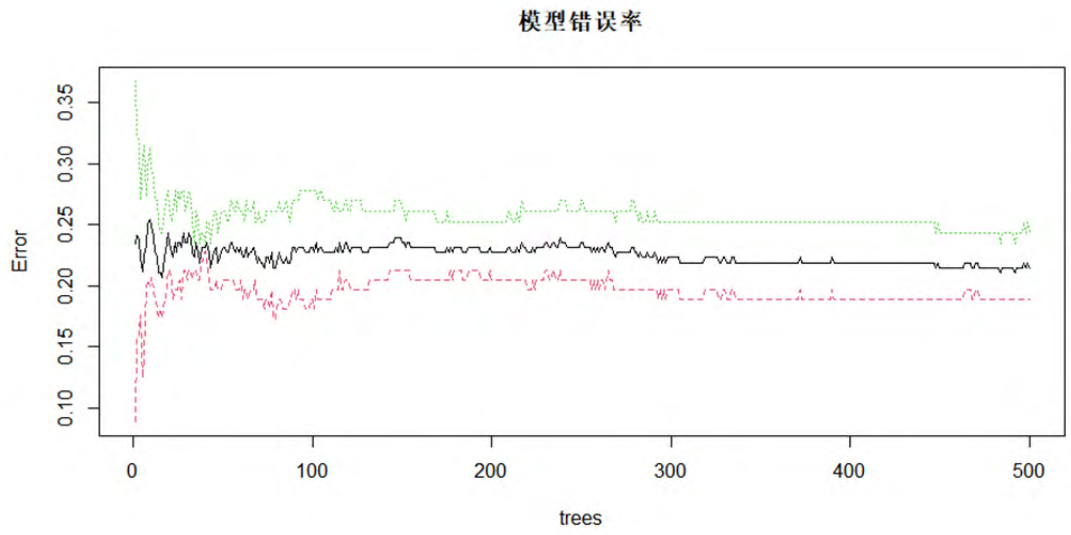

图 6　模型错误率

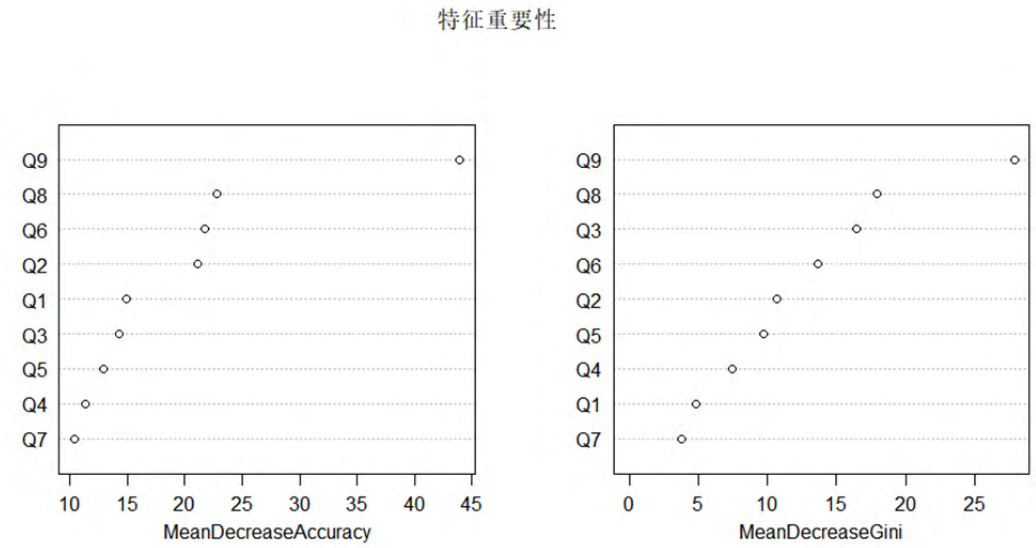

图 7　基于准确率和基尼系数的特征重要性

利用 MDS 算法，衡量各个样本的曼哈顿距离，对于数据进行降维便于更加简单地观察数据的分布。

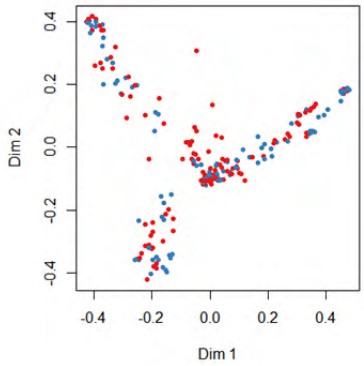

图 8　基于MDS算法的数据降维

为了对模型的分类效果进行检验，我们对于模型的 AUC 分数进行计算。分类器的 AUC 值等价于将随机选择的正样本排序在随机选择的负样本之前大概率，在机器学习中常常用 AUC 判断分类器优劣的标准。

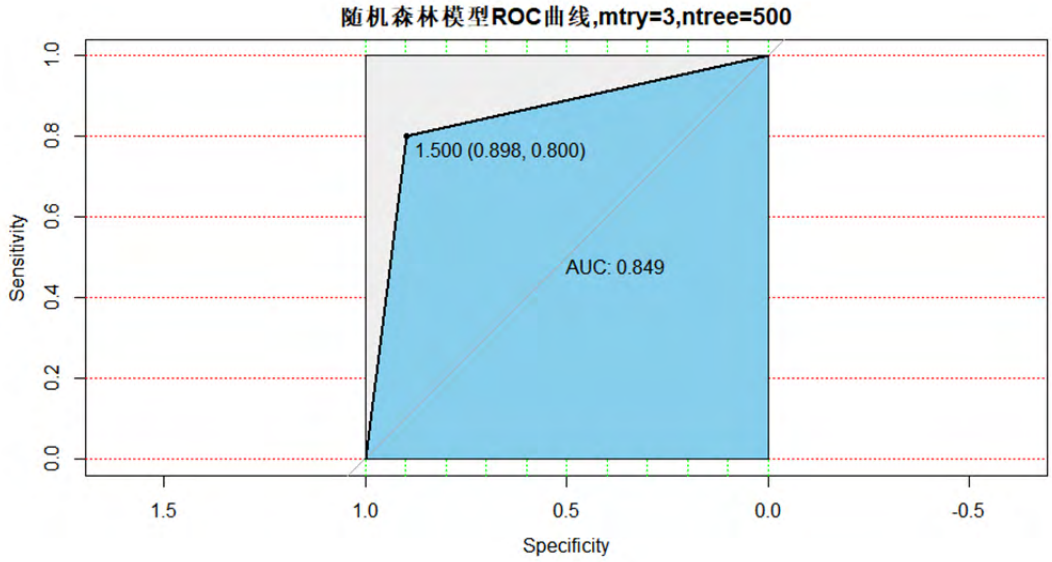

图 9　随机森林模型的ROC曲线

模型的 AUC 值达到 0.849，效果较好，但是为了提高模型的效率，我们根据不同数量决策树的错误率的结果，决定将错误率较少，且再增加决策树也没有提高正确率的决策树的数量作为模型最后的参数为 200。再次计算模型中每个变量的特征重要性和 ROC 曲线。

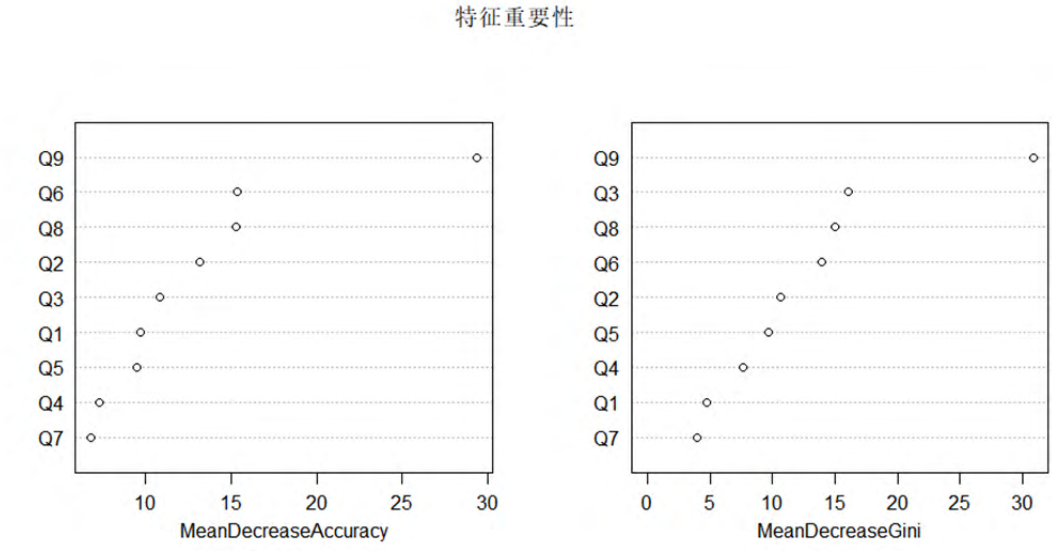

图 10　改进后模型的特征重要性

根据图 3.2.5 和图 3.2.7 的结果，我们发现模型的 AUC 值从 0.849 增加到 0.858，随机森林模型的分类器效果提高，且模型计算量减少。

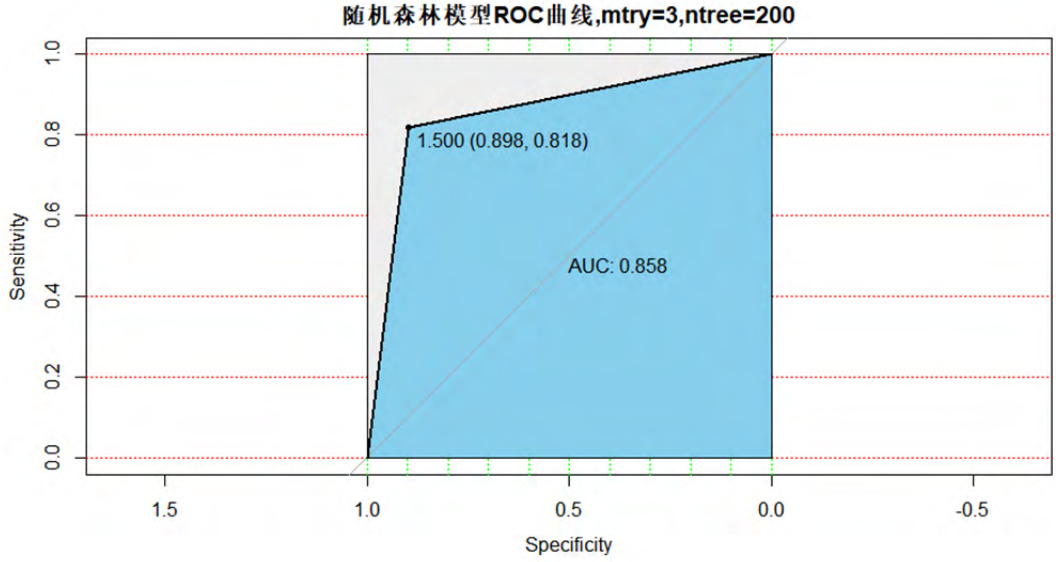

图 11　改进后模型的ROC曲线

4.模型结果分析-随机森林模型变量相对重要性分析

我们发现 Q9 的特征重要性远大于其他变量，说明家长如果给孩子报名其他类型的培训班也会有更强烈的意愿给孩子报名少儿编程培训课程，主要是由于这类家长对于孩子的学习非常重视且愿意为了孩子的教育付出一定的经济成本，在

信息教育的大背景下，编程在教育中的地位越来越高，这类家长给孩子报名少儿编程培训课程的意愿也就越强。因此，我们建议少儿编程培训机构可以与其他培训机构进行合作，进行相互的宣传，互相投放广告，将宣传目标瞄向潜在消费群体，提高宣传效果。

（三）主成分分析模型

1.模型的介绍与建立

在研究多变量问题时，变量过多会增加研究的复杂性，并且在一定情况下，变量之间存在相关联系，其所反映的信息存在一定重叠。主成分分析即将多个存在一定相关关系的变量转化为新变量，将重复的变量删除并保证新变量所反映出的信息尽可能的保持原有信息，从而简化数据结构。

在对少儿编程教育培训机构的研究中，家长对培训机构的重视层面研究方向较少，对培训机构的重视层面反之决定了培训机构的发展方向，因此，有必要建立客观研究家长对培训机构的重视层面的方法。在研究家长重视层面时，相关随机变量存在部分重叠，因此，本论文采用了主成分分析法研究家长对培训机构的重视层面。

主成分分析法研究方法：指标数据标准化；指标之间的相关性判定；确定主成分个数 m；确定主成分表达式。

① 原始指标数据的标准化采集 p 维随机变量

$$\mathrm{x} = (x_1, x_2, x_3, \ldots\ldots x_p)^T \tag{3}$$

n 个样品

$$x_j = \left(x_{j1}, x_{j2}, \ldots\ldots x_{jp}\right)^T,\ \mathrm{j} = 1,\ 2,\ \ldots\ldots \mathrm{n}, \mathrm{n} > \mathrm{p} \tag{4}$$

构造样本阵，对样本阵元进行如下标准化变换：

$$Z_{ij} = \frac{x_{ij} - \bar{x}_j}{S_j},\ \mathrm{i} = 1,2,\ \ldots\ldots \mathrm{n}; \mathrm{j} = 1,2, \ldots\ldots,\ \mathrm{p} \tag{5}$$

其中

$$\bar{x}_j = \frac{\sum_{i=1}^{n} x_{ij}}{n} \tag{6}$$

$$s_j^2 = \frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n-1} \tag{7}$$

得标准化阵 Z；对标准化阵 Z 求相关系数矩阵

$$\mathrm{R} = \left[r_{ij}\right]_p xp = \frac{Z^T Z}{n-1} \tag{8}$$

其中

$$r_{ij} = \frac{\sum z_{kj} \cdot z_{kj}}{n-1}, \mathrm{i,j} = 1,\ 2,\ \ldots\ldots,\ \mathrm{p} \tag{9}$$

② 解样本相关矩阵 R 的特征方程

$$\left|\mathrm{R} - \lambda I_p\right| = 0 \tag{10}$$

得 p 个特征根，确定主成分按

$$\frac{\sum_{j=1}^{m} \lambda_j}{\sum_{j=1}^{p} \lambda_j} \geq 0.85 \tag{11}$$

确定 m 值，使信息的达 85%以上，对每个λ_j，j=1，2，……，m，解方程组$\mathrm{Rb} = \lambda_j$，b 的单位特征向量b_j^0；

③ 将标准化后的指标变量转换为主成分

$$U_{ij} = z_i^T b_j^0,\ \mathrm{j} = 1,\ 2,\ \ldots\ldots,\ \mathrm{m} \tag{12}$$

④ U_1称为第一主成分，U_2称为第二主成分，……，U_p称为第 p 主成分；

对 m 个主成分进行综合评价。对 m 个主成分进行加权求和，即得最终评价值，权数为每个主成分的方差贡献率。

2.模型的求解与结果

进行 KMO 和 Bartlett 检验，可得， KMO 统计量为 0.816，变量相关性较强适于因子分析，并且 Bartlett 球形检验对应的 P 值为 0.000，说明符合标准，各变量在一定程度上相互独立。

W3

表 8　信效度检验表

Cronbach ’s Alpha		0.735
取样足够度的 Kaiser-Meyer-Olkin 度量		0.816
Bartlett 球形检验	近似卡方	1828.731
	df	36
	Sig.	0.000

表 9　总方差解释

	初始特征值			提取载荷平方和		
成分	总计	方差百分比	累积 %	总计	方差百分比	累积 %
1	2.280	25.333	25.333	2.280	25.333	25.333
2	1.597	17.743	43.076	1.597	17.743	43.076
3	1.224	13.598	56.673	1.224	13.598	56.673
4	1.062	11.795	68.468	1.062	11.795	68.468
5	.916	10.176	78.644			
6	.792	8.804	87.448			
7	.593	6.584	94.032			
8	.533	5.919	99.952			
9	.004	.048	100.000			

由图可以看出前四个主成分特征值均大于 1，故提取前四个主成分，第一、第二、第三和第四主成分已经累积达到 68.468%，可以代表大部分数据的信息

表 10　成分得分系数

	成分			
	1	2	3	4
硬件设施	-.169	.058	-.496	.508
课程费用	-.303	-.183	.261	-.146

离家距离	-.277	.208	.229	-.304
教学质量	.615	.385	-.374	-.350
教学氛围	.208	.373	-.294	-.043
教师质量	.269	-.302	-.006	-.142
课程体系	.205	-.176	.338	.640
上课方式	.259	.093	.326	.021
服务态度	.256	.357	.062	-.124

由表可知，大多数家长较为关注教学质量和教师质量，同时上课方式也是多数家长选择培训机构的主要原因，在选择培训机构时家长并不关注硬件设施、课程费用和离家距离。

因此我们建议培训机构可以将注意力集中在提高实际教学质量上，并且可以通过开展线上课程缓解引入老师的资金问题，在提高教学质量的同时也可以保证培训机构的成本得到控制。在选拔教师时，主要面向于专业与计算机相关的的毕业生，并且应着重考察教师的实际编程能力，重视基础知识的完备性。同时可以设定奖励机制，让老师有更多的教学动力，并且根据不同学生的需求设定不同教学方案，满足不同需求，制定不同定价策略。

四、网络数据获取与文本挖掘模型分析

（一）数据收集及预处理

1.数据爬取

为了详细了解消费者在少儿编程培训机构的真实体验与感受，我们利用Python编写网页爬虫自动爬取在美团，大众点评等生活软件上西安市少儿编程培训机构消费者最近一年的相关评论。部分消费者的评论内容较为单一，只有简单的“好评”，“好”，“非常棒”等此词，不足以判断消费者的具体感受，因此

对这些评论不进行爬取。总共爬取相关有效评论346条。

2.数据预处理

问题数据可以对数据的完整性和合理性构成影响，从而影响数据剖析结果。因而，在进行文本主题挖掘和情感分析之前，非常关键的步骤是对采集到的文本进行预处理。本文以每条评论文本为单位进行处理，由于在评论中经常会出现相同的词，这些词表达的意义基本相同，比如：“好好好”“喜欢喜欢”之类的评价语句，如果直接使用这些语句进行文本挖掘，会对实验的可靠性造成很大影响。因此，在对数据进行预处理时首先要对评论文本去重。常见的文本去重方法大多以计算文本之间的相似度为基础。由于此调查中的文本数据为短文本，重复文本大都是相同的词语或单个字，因此采用机械压缩的方法，将完全重复的文本压缩成单个词语或字。

（二）词频分析

我们利用Python的jieba库对所爬取的评论进行分词，把所有评论分解成中文的词语，并进行词频统计。再使用wordcloud库将词频较高的词语画成词云图，更加直观地展示消费者的实际感受和需求，从而更好地为少儿编程培训机构提出相关改进建议。

我们发现“老师”、“孩子”、“课程”、“编程”等词的热度较高，大多数消费者偏好老师对于课程内容能够讲解的清晰”、“通俗易懂”、“简单”，便于孩子能够更好地理解和接受少儿编程的培训内容。因为编程课程对于大多数孩子而言，难度可能较大，需要大量时间去理解和接受课程内容，所以消费者也希望老师在课堂上能够做到“耐心”、“细致”，在编程方面能够有较为专业和全面的掌握。同时，我们根据所爬取的词语建立词义网络图分析。

我们发现，在课程设置上，消费者更加偏好于老师能够设置更多的解决学生疑难问题的环节。我们发现大多数消费者给孩子报名少儿编程培训班的主要目的为：提高孩子的逻辑思维、激发孩子对于编程的兴趣、让孩子掌握一定的编程能力。而大多数消费者了解少儿编程培训机构的途经是通过朋友的推荐和介绍，因此课程质量和孩子的上课体验是培训机构能否吸引更多消费者的关键。

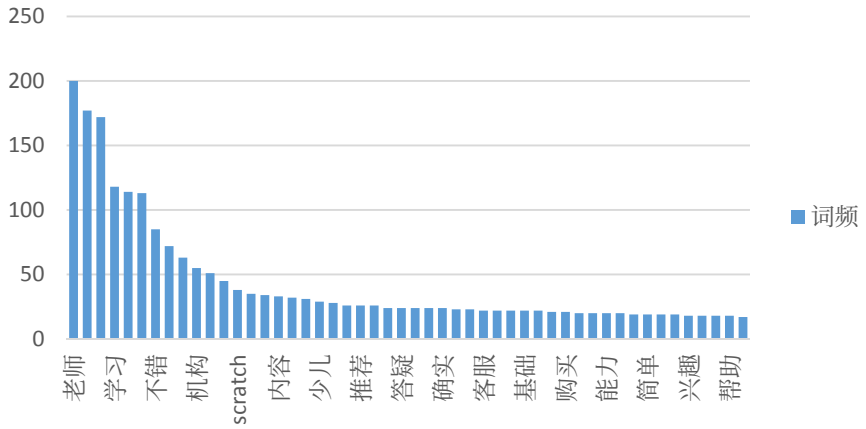

图 12　词频分析频数表

图 13　词云图

2020年因为疫情的原因线上课堂在社会逐渐变得流行，而编程也是一项实践性非常强的课程，因此我们特别关注到，部分开设线上课程的培训机构的消费者

认为线上课程具有价格低，可重复观看的特点，孩子能够在课后反复观看老师所讲授的课程内容，从而在课后进行编程训练，相比线下教学，孩子可能会有更多的时间去训练。同时，家长也较为关注少儿编程培训的相关费用，线上教学在不降低课程质量的前提下可以极大地减少培训机构的场地费用、硬件设施费用、教师培训费用，从而可降低课程价格，提高价格竞争力。

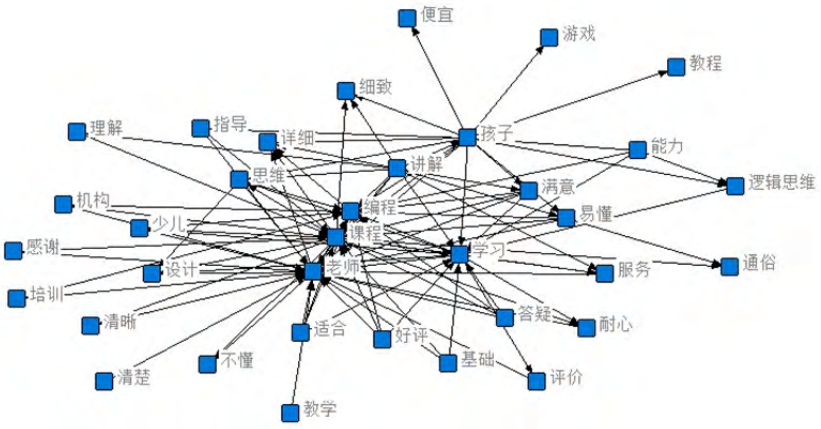

图 14　网络爬取的文本数据词意网络图分析（1）

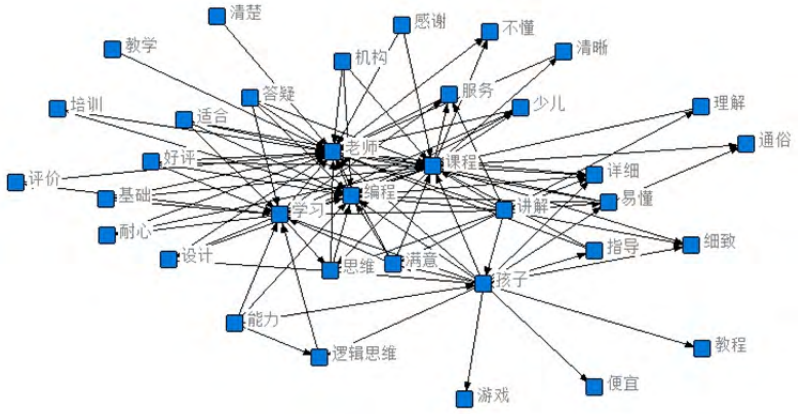

图 15　网络爬取的文本数据的词意网络图分析（2）

五、少儿编程培训机构的发展策略建议

（一）宣传方式的建议

通过问卷调查和数据分析，我们发现单一的以广告为宣传手段，并不能够给消费者带来较深刻的影响。而如果能够让消费者成为宣传方式即提高课程的质量从而让自己的名声变好，报名过该课程的消费者也会把本课程推荐给身边的朋友，达到宣传的手段。同时我们发现培训机构应该在宣传课程质量的同时，注意普及编程在升学方面的政策，让更多的家长了解编程在未来教育中的地位，把握市场先机。

根据多元Logistic回归模型和随机森林模型的分析结果，我们发现报名过其他培训班的受访者也会更加愿意给自己的孩子报名少儿编程课程，因此少儿编程培训机构可与其他机构合作，比如在“奥数”、“少儿英语”等培训班适当投放少儿编程的广告，将适量的广告集中于消费意愿更强的家长，增加宣传效果同时较少投放成本。

（二）课程设置的建议

建议培训机构在开始教学前对于所有学生进行编程能力测试，根据测试成绩反映学生的编程基础，并依此为依据进行分组教学。在编程基础较弱的组别，大多数家长给孩子报名少儿编程培训课程的主要目的是为了激发孩子的兴趣，提高孩子的逻辑思维能力，所以培训机构在课程设置上应该注意不能单纯地以应试教育的形式进行教学，更多的是需要开放式地讨论，增加老师与孩子的互动，让孩子能够进行自我思考，有疑问时可以得到及时解答。

在编程能力较强的组别中，家长更多是想要孩子能够获得专业的编程教育，进一步提高编程能力，在竞赛中能够斩获一定的成绩，因此培训机构对于这部分的教学应该以专业培训为主，适当地布置相应作业。

（三）教师选拔的建议

教师是课程质量的保证，根据问卷调查结果，我们建议培训机构在选择教师时应该更加注重教师的实际编程能力，而不是一味地追求“名校”学历。因为编程是一个实践性非常强的学科，所以培训机构也需要注意教师的耐心是非常重要的，每个学生在实际编程过程中的学习能力是不相同的，教师需要有足够的耐心去解决学生的问题。由此出发提出以下解决措施：第一，在每节课之后设置反馈环节，让学生评价老师的上课情况，并对教师的教学态度做出点评；第二，定期进行教师考核制度，督促教师也能够不断学习新的知识更好地开展教学；第三，实施教学激励措施，比如在以成绩为主的培训班中给予培训教师一定的激励手段，提高竞赛成绩。

（四）开展线上教学的建议

目前西安市的少儿编程培训机构开展线上教学的情况并不多，但是绝大多数的受访者都希望能够将线上教学与线下教学结合起来。因为线上教学可以节省孩子大量花费在路上的时间，同时也可以保证孩子的安全。当然相比线下教学，线上教学最主要的问题就是教师与同学的互动不足，老师难以实时掌握孩子的学习情况，因此我们提出以下解决措施：第一，仿照MOOC形式建立交互式平台，使得老师和学生的交流更加方便，学生也可以有更多时间和更加便捷的方式去进行编程实践；第二，引入人脸识别系统，可以在线上教学时也能够较为准确地观察到学生的上课情况，将线上教学的缺点尽可能减小；第三，通过布置线上编程实践作业的方式，让老师快速知晓学生的疑难点问题，相比线下教学更为方便。

同时，也可以聘请专业能力较强的老师通过录制微课的方式，让学生能够自我掌握上课时间，并且培训机构也可以将微课作为售后产品，弥补课程教学的不足之处，提高课程质量。

致谢

本论文是在XX大学XX学院的XXX老师的亲切关怀和悉心教导下完成的。XXX老师对我们的论文提出了很多具有建设性的意见，在初步调查时，给予我们关于调查方向和调查重点的建议。在建立模型和论文撰写时，提出很多重要的思路，帮助我们解决问题，在此谨向XXX老师以诚挚的谢意和崇高的敬意。

参考文献

[1]. 中国少儿编程行业研究报告 2018年[A].上海艾瑞市场咨询有限公司,艾瑞咨询系列研究报告[C].:上海艾瑞市场咨询有限公司, 2018.

[2]. 2017-2023年中国少儿编程市场分析预测及发展趋势研究报告[A].北京智研咨询有限公司,智研咨询系列研究报告[C]. :北京智研咨询有限公司2017.

[3]南方日报.2019年全国教育事业发展统计公报:全国在校生2.82亿人.https://m.sohu.com/a/396502335_161795.2020年5月20日发布.

[4]国家统计局.中国统计年鉴[J].北京:中国统计出版社. 2020.

[5]陕西统计局,国家统计局陕西调查总队.陕西统计年鉴[J].北京:中国统计出版社.2020.

[6]吴明隆.问卷统计分析实务——SPSS操作与应用[M].重庆:重庆大学出版社,2010.

附录 调查问卷

关于西安市少儿编程市场消费者需求偏好的调查问卷

您好，我们是XXX大学的调查员，正在做一份关于西安市少儿编程市场消费者需求偏好的调查。诚挚地邀请您参与此次的问卷调查，本次调查数据仅作科研之用，问卷采用匿名制，我们对您的信息进行严格保密，请您根据自己的实际情况填写，希望得到您真实的想法与宝贵的意见，非常感谢您的支持与配合。

1.您的性别？

□男

□女

2.您的年龄？

□30岁以下

□31—35岁

□36—40岁

□41—45岁

□46—50岁

□51岁以上

3.您的家庭住址所在地区？

□新城区

□碑林区

□莲湖区

□灞桥区

□未央区

□雁塔区

□阎良区

□临潼区

□长安区

□高陵区

□鄠邑区

4.您的受教育程度？

□未上过学

□小学

□初中

□高中

□中专

□大学专科

□大学本科

□研究生及研究生以上

5.您的月收入情况？

□3000元及以下

□3001—5000元

□5001—7000元

□7001—9000元

□9001元以上

6.您的职业为？

□以理工类工作为主

□以社会科学类工作为主

□以艺术类工作为主

□自由职业

□其他

7.您的家人中有从事IT行业的吗？

□有

□无

8.您的孩子的受教育程度？

□无

□学龄前

□小学1-3年级

□小学4-6年级

□初中

□高中

9.您的孩子是否在上补习班？

□是

□否

10.您的孩子是否接受过少儿编程培训？

□是

□否

11.您可以接受孩子每年课外培训的费用？

□3000元以下

□3001—5000元

□5001—7000元

□7001—9000元

□9001元以上

12.您可以接受孩子每年学习少儿编程的费用？

□3000元以下

□3001—5000元

□5001—7000元

□7001—9000元

□9001元以上

13.您通过以下哪些方式了解少儿编程教育？（多选）

□朋友介绍

□广告

□查阅相关资料

□机构工作人员介绍

□学校老师的介绍

14．请您给您认为印象深刻的宣传方式进行评分

	1	2	3	4	5
朋友介绍和推荐					
培训机构的相关广告					
查阅相关资料					
机构工作人员的介绍					
学校老师的介绍					

15.您对编程的了解程度大致为？

□很熟悉

□基本熟悉

□一般了解

□不太了解

□不清楚

16.您对少儿编程在升学、竞赛及其他方面的相关政策的了解程度？

□很熟悉

□基本熟悉

□一般了解

□不太了解

□不清楚

17.请您对以下几个让孩子接受少儿编程的理由中进行排序？

□受周围环境影响

□激发兴趣

□提高编程能力，为后续学习作铺垫

□避免孩子沉迷游戏

□参加编程类赛事，让孩子在未来升学的过程中占优势

□提高孩子的逻辑思维、观察力和创造力等

□孩子喜欢就好，没有目的

□其他

18.请您对培训机构各个方面的重要程度进行评分

	1	2	3	4	5
硬件设施					
课程费用					
离家距离					
教学质量					
教学氛围					
教师质量					
课程体系					
上课方式					
服务态度					

19.您希望孩子上培训课的班级人数为？

□一对一或一对二

□3—5人

□6—10人

□11—30人

□31人以上

20.您是否希望培训机构能够进行分组教学？

□是

□否

21.您希望培训机构按照什么标准进行分组？

□年龄

□编程能力

□学习能力

□定期考试制度

22．请您对老师在以下方面的重要程度进行评分

	1	2	3	4	5
老师指导学生在竞赛中所获得过的成绩					
老师的影响力					
老师自己曾获得的荣誉					

老师的教学态度					
老师的学历					
老师的专业能力					

23.您希望老师与孩子课堂的课堂互动程度？

□不互动

□一般互动

□经常互动

□频繁互动

24.您希望老师对学生的监管力度？

□宽松

□比较宽松

□比较严苛

□非常严苛

25.您是否希望老师布置课后作业？

□是

□否

26.您希望少儿编程培训机构每节课的时间安排为？

□30min以内

□31—60min

□61—90min

□91min以上

27.您希望孩子一周的编程课程的总时长为？

□30min以内

□31—60min

□61—90min

□91min以上

28.您可接受的上课地点距家车程（公共交通）？

□20min内

□40min内

□60min内

□61min以上

29.线上教学和线下教学相比，您更加接受哪种方式？

□线上教学

□线下教学

□线上教学与线下教学相结合

（29选择“线上教学”）30.您选择线上教学的原因是？（多选）

□能够节省孩子的时间

□相较于线上课程而言，价格较低

□孩子的安全能够得到保证

□有更多实际操作的时间

□上课时间可以自由选择

（29选择“线下教学”）31.您选择线下教学的原因是？（多选）

□课程质量和效果更好

□老师能够监督孩子的学习

□孩子和老师能够有更多交流

□培训机构离家较近，所花时间不多

□老师和孩子的交流会更多

32.您对于目前西安市少儿编程培训机构的改进建议？