Management Science and Engineering 管理科学与工程, 2018, 7(1), 48-54
Published Online March 2018 in Hans. http://www.hanspub.org/journal/mse
https://doi.org/10.12677/mse.2018.71006

Hans汉斯

Prediction of Conversion Rate of E-Commerce Orders Based on Data Mining

Zhi Liu

School of Economics and Management, Beijing Jiaotong University, Beijing
Email: lzzy201311@163.com

Received: Mar. 6th, 2018; accepted: Mar. 20th, 2018; published: Mar. 29th, 2018

Abstract

The growth of e-commerce business tended to be flat and the ultra-high-speed growth brought by traffic and mobile Internet dividend was basically over. E-commerce business entered a refined operation phase. Different from the offline operation mode, the electricity supplier's order quantity is mainly affected by the marketing strategy and can collect detailed operational data, which has great potential for utilization. In this paper, by way of modeling, linear regression model and non-linear model, that is machine learning model, are used to predict the conversion rate. The purpose of this paper is to help platform-based e-commerce providers to predict the conversion rate and to assist in setting different conversion rates according to the different stages of business objectives and allocating marketing resources. At present, most researches mainly focus on theoretical research, lacking of analysis of specific application links. From the perspective of data and models, this paper starts from the feature selection to the selection of models as the logic to carry out modeling analysis of the impact of marketing strategy and draws corresponding conclusions, with practical guidance value.

Keywords

E-Commerce, Order Conversion Rate, Models, Forecast, Machine Learning

基于数据挖掘的电商订单转化率的预测

刘　治

北京交通大学经济管理学院，北京
Email: lzzy201311@163.com

收稿日期：2018年3月6日；录用日期：2018年3月20日；发布日期：2018年3月29日

文章引用：刘治. 基于数据挖掘的电商订单转化率的预测[J]. 管理科学与工程, 2018, 7(1): 48-54.
DOI: 10.12677/mse.2018.71006

摘　要

电商营业增速趋于平缓，流量和移动互联网红利等带来的超高速增长基本结束，电商业务进入精细化运营阶段。不同于线下的运营模式，电商的订单量主要受营销策略的影响，且可以收集到详细的运营数据，具有极大的利用潜力。本文通过建模的方式，利用线性模型和非线性模型，尤其是机器学习模型，以转化率为目标进行回归预测。本文的研究意义及目的在于帮助平台型电商进行转化率的预测，以辅助根据不同阶段的企业目标制定不同的转化率目标，进行营销资源的分配。目前的研究多以理论研究为主，缺少对具体应用环节的分析，本文从数据和模型的角度，从特征选择到选择模型为逻辑进行展开，对营销策略的影响进行建模分析，并得出相应的结论，具有实践指导价值。

关键词

电子商务，订单转化率，模型，预测，机器学习

1. 引言

2016 年以来，京东、阿里营收同比增速下降到 40%到 50%，相比过去三位数的增速，这标志着电商增速的放缓。电商开始进入已有业务的精耕细作和新业务拓展的新阶段。已有业务的精耕细作主要是营销方式的提升和供应链、物流和配送等后端服务能力的优化，利用电商数据价值赋能整个价值链。新的业务主要是 O2O (线上线下整合，以盒马生鲜为代表)、跨境电商、农村电商。本文聚焦于电商数据价值赋能领域，基于统计学模型和数据挖掘模型，研究电商营销方式和流量对转化率的影响。但目前对于电商订单转化率的研究还很少，尤其是用机器学习模型建模的方法。

消费者在电商平台的消费行为可以分为以下几个阶段，需求产生、搜索(浏览)信息，评估选择、购买决策和评价反馈，到下一次新的消费行为产生一个新的周期[1]。因为本文的研究对象是转化率，着重分析的是搜索(浏览信息) - 评估选择环节 - 购买决策阶段。

消费者在电商平台的购物行为不同于在线下零售店，具有信息量大、不可接触实物、买家与卖家沟通较少的特点，极大受电商平台的营销行为的影响。目前的研究中讨论了营销的长期因素和短期因素。刘贵容等分析了影响电商转化率的 9 个因素及相关关系，归纳为 4 组相关因素，有消费能力、访问目的和网购体验和其他[2]。韩睿等通过对买赠、返券、打折三种不同方式对消费者价值感知和购买行为的实证研究，发现三种不同方式会对最终购买行为产生不同的影响[3]。此外对于电商转化率的研究还有关于精准营销、用户界面、IT 系统、搜索、基于某一商家型电商等影响转化率的细分领域的研究。还有一些新的商业策略也值得关注，如会员制、供应链、O2O 等新的商业模式，都会影响到转化率。

总的来说影响订单转化率的原因是用户体验、促销策略和产品，本文将促销方式和浏览量作为内生变量，对订单转化率进行回归建模。经过研究以浏览量和促销费用作为预测变量建立的模型平均绝对误差率达到 91%，具有较高的实用价值。

2. 理论模型

发现预测变量和解释变量之间规律的问题分为分类问题和回归问题，从知识发现(Knowledge Dis-

covery in Database)的角度，预测变量是数值型变量即为回归问题，预测变量若为类别型变量则为分类问题。回归问题可以根据预测变量与解释变量是否为线性关系，可以分为线性回归模型和非线性回归。线性回归的优势在于简单和易于描述，但因为现实生活中的事物间很少符合严格的线性关系，导致在现实应用中往往拟合效果不及非线性模型。此外对异常值敏感和多重共线性等问题也常常降低了线性回归模型的拟合效果。非线性模型中机器学习算法，在当前企业积累大量数据的环境下应用变的可能。种种优势得到突显，可以很大程度上基于数据给出客观的模型，而较少受到主管判断的干扰。且在各种模型拟合效果的检验中表现远胜过线性模型。本文通过探索不同模型的回归效果，选择适合本文描述情境下最优的模型。

2.1. 多元线性回归模型

线性回归是利用梳理统计中回归原理来确定预测变量与解释变量之间的关系，并且假设预测变量与每个解释变量之间为线性关系。多元线性回归模型是解释变量为两个或两个以上的情况下的回归，为了得到最优的拟合效果，线性回归在寻找模型最优参数时常用的损失函数是最小二乘法(OLS)。多元线性回归原理同一元线性回归相同，但计算上复杂得多，因此需要计算机来完成。

设预测变量为 Y，k 个解释变量分别为 $X_1, X_2, \cdots, X_k$，描述 Y 与 $X_1, X_2, \cdots, X_k$ 之间线性关系的方程成为多元回归模型，其一般形式表示为：

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_k X_k + e \tag{1}$$

式中 b_0 为常数项，$b_1, b_2, \cdots, b_k$，为回归系数，e 为误差项。

式(1)中误差项 e 反映了除模型中给出的解释变量 $b_1, b_2, \cdots, b_k$ 与预测变量 Y 的线性关系之外的随机因素对 Y 的影响，是不能由 $b_1, b_2, \cdots, b_k$ 与 Y 之间的线性关系解释的变异性。建立多元线性回归模型时，为保证模型具有较好的拟合和预测效果，需要一下四个基本假设：

1) 误差项 e 的期望值为 0；
2) 对解释变量所有样本观察值的随机误差项 e 都独立同分布，且为正态分布；
3) 解释变量是确定性变量，不是随机变量，与随机误差项彼此之间相互独立；
4) 解释变量之间不存在精确的线性关系，即解释变量的样本观测值矩阵是满秩矩阵。

2.2. 非线性回归模型

本文对于非线性回归模型只关注机器学习模型，非线性回归模型主要分两种，一种是将非线性关系映射到线性空间(如支持向量机、神经网络)，另外一种是直接用具有非线性拟合能力的模型(如决策树、随机森林)。

随机森林(Random Forest)是一种基于分类(回归)决策树的组合预测模型是一种较新机器学习模型，2001 年由 Breiman 提出。随机森林弥补了决策树过度拟合的缺点，而且对缺失值和异常值不敏感，具有更强的泛化能力。随机森林中每棵树的训练需要训练样本集和特征变量两个维度的数据，其中每棵树所使用的训练样本集是从总的训练样本中，从观察对象和特征两个维度，有放回的采集出来。总的训练集中的有些样本可能多次出现在一棵树的训练样本集中，也可能从未出现在一颗树的训练样本中。每棵树所使用的特征是按照设定的数量或比例从所有特征中无放回随机抽取的。在每一棵决策树的建立过程中，包含采样与完全分裂两个子过程：1) 随机采样过程，随机森林对训练样本集数据要进行和列的采样。其中行采样采用有放回的方式，列采样是从 M 个特征中无放回抽取 m 个(m<<M)特征；2) 以采样得到的数据为根节点数据，通过完全分裂的方式建立决策树，并继续向下分裂，直到所有样本都是指向同一个分

类[4]。成为叶节点，预测值为叶节点目标变量的加权均值。

支持向量机(Support Vector Machine)是在统计学理论(statistical learning theory, SLT)基础上发展起来的一种数据挖掘方法，1992 年由 Boser，Guyon 和 Vapnik 提出。在解决小样本，非线性及高维的分类和回归问题中表现出很多优点。支持向量机用于研究输入变量与数值型输出变量的关系的回归应用，简称为支持向量机回归(Support Vector Regression)。对于非线性的支持向量机回归，通过一个非线性映射(即核函数)把样本非线性映射到高维特征空间，然后对这个空间进行线性回归。相比线性回归中通常采用的最小二乘法对每个观测的误差计入损失函数，支持向量机采用ε—不敏感损失函数，误差函数数值小于指定值 ε 的观察给损失函数带来的“损失”将被忽略。而且支持向量机形式上类似一个神经网络，输出是中间节点的线性组合，每个中间节点对应一个支持向量[5]。

人工神经网络(Artificial Neural Network, ANN)是一种模拟人脑思维的计算机建模方法，20 世纪 40 年代心理学家 McCulloch 和数学家 Pitts.W.H 建立了著名的阈值加权和模型(M-P 模型)，标志这神经网络研究的开始。在人工神经网络的发展过程中，对生物神经系统已经从不同层次的描述和模拟，提出了各种各样的神经网络模型，其中具有代表性的网络模型有：感知神经网络、线神经网络、BP 神经网络、径向基函数网络、自组织网络、反馈网络等。目前在实际应用中，大多数人工神经网络模型采用的是前馈反向传播网络(Back-Propagation-Network 简称 BP 神经网络)或者它的变化形式。标准的 BP 网络是根据 W-H 学习规则，采用梯度-F 降算法，对非线性可微分函数进行权值训练的多层网络。并且理论已经证明，三层 BP 神经网络只要隐节点数足够多，理论上就可以模拟任何复杂的非线性映射。对于回归问题，隐层的一个节点就是一个回归平面，人工神经网络的训练过程就是通过对训练集的学习寻找最佳回归平面的过程。

本论文就采用三层 BP 神经网络，以某电商业务基本成熟后一年的销售数据进行建模分析，以订单转化率为预测变量。

3. 实证分析

D 公司是一家国内知名跨境电商公司，所使用的数据是 2016 年 8 月到 2017 年 7 月间一年的数据，数据颗粒细分到每天。

3.1. 数据描述

本文的研究对象是客户满意度、直降、满减和优惠券的营销组合对订单转化率的影响，所以预测变量就是订单转化率，特征变量分别用销售当天的独立访客量、直降金额、满减金额和优惠券来代表。图 1 和图 2 是对这四个数据的直观展示。考虑到该场景下数据的异常是由业务部门的促销活动导致，即异常值正常反应了业务运营结果。

图 1 可以看到 D 公司的转化率数据近一年来呈现总体平稳的趋势，没有显著的上升或下降，有几处显著的高峰和低谷。峰值的时间符合常识，大多是都是平台进行了促销活动或者节日。通过对转化率分布的探索发现，以双边 5%为异常值上下限，共有 27 个异常值，其中 16 个为异常高，11 个异常低中有 7 天处于春节期间。仅有 5 天为未守节日影响的转化率异常，日期分别为 2016 年 7 月 20 日、2016 年 9 月 21 日、2017 年 1 月 11 日和 2017 年 3 月 1 日和 2017 年 5 月 12 日。

独立访客和订单转化率的关系较为复杂，独立访客既是客户对电商平台满意度衡量的一个指标，但独立访客是一个与订单转化率相独立的变量，独立地受企业营销策略和产品的影响。在图 2 中我们也可以看到这一规律，大多数时候独立访客和订单转化率同时出现波动，但有些时候波动时间不一致，且波峰的变化幅度也与订单转化率不同。

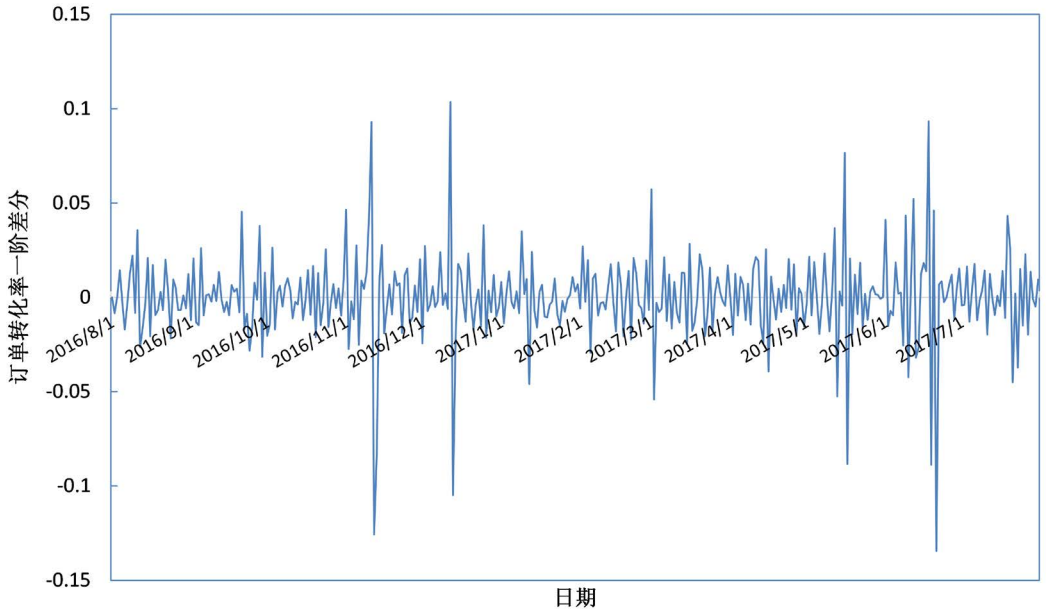

Figure 1. Conversion rate and different promotional investment line chart
图 1. 转化率与不同促销投入折线图

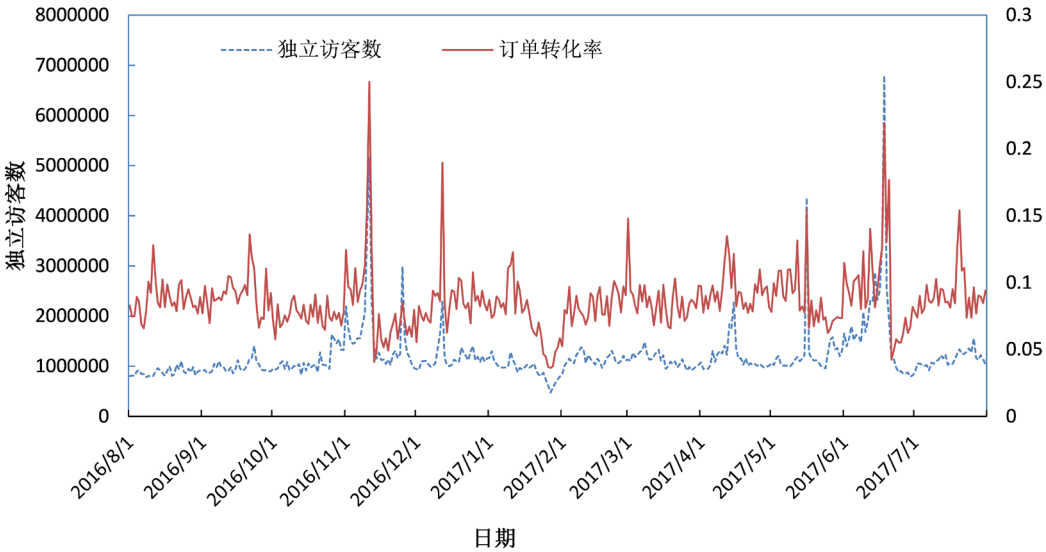

Figure 2. Order conversion rate and independent visitors line chart
图 2. 订单转化率与独立访客折线图

3.2. 建模

在多要素所构成的系统中，当研究某一个要素对另一个要素的影响或相关程度时，把其它要素的影响视作常数(保持不变)，即暂时不考虑其他要素影响，单独研究两个要素之间的相互关系的密切程度，所得数值结果为偏相关系数[6]。使用 SPSS Statistics 22 计算各变量间偏相关系数如表 1，表明各变量间存在线性相关性，但相关性不是很强。支持进行多元线性回归的建模。

R 是一种用于统计计算和画图的编程语言和开发环境，其提供了丰富的统计(线性和非线性建模，经典统计学检验，时间序列分析、分类，聚类等)和画图技术，而且具有极高的拓展性。R 语言源于经典的

S 语言，S 语言通常是统计方法研究的首选工具，R 提供了开源的途径来参与该活动。

使用 R 语言中的 e1071、randomforest 等工具包建立模型和预测，具体的多元线性回归使用函数 lm()，随机森林使用 randomForest()，支持向量机使用 svm()，BP 神经网络使用 nnet()。参数选择使用网格法的搜索方法，以训练集合的 MES 值为择优标准选择最佳的模型参数。

3.3. 模型比较

为对多元回归模型、随机森林模型、支持向量机模型和神经网络模型的模型拟合能力和预测能力进行比较分析。本文使用的模型方法是，首先随机划分模型的训练集和测试集，结合使用数学指标来反应回归效果。综合对训练集和测试集的参数比较结果，选择最优的回归模型。本文使用平均误差平方和(MES)和绝对误差百分率(MAPE)这两个指标进行比较。实验中通过 30 个不同的随机数种子进行 9:1 的训练集与测试集的划分，求得不同模型的两个参数，然后进行平均得到表中的参数。

$$\mathrm{MSE}=\frac{1}{n}\sum\left(YS_i-Y_i\right)^2$$

$$\mathrm{MAPE}=\frac{1}{n}\sum\left|\left(YS_i-Y_i\right)/Y_i\right|\times 100\%$$

从表 2 可见，无论是你和能力还是预测能力，非线性模型中支持向量机的拟合效果最佳。线性回归模型拟合效果和预测效果都是最差。神经网络模型拟合效果最好，但过度拟合严重。随机森林对训练集的拟合最好，但是在测试集上表现较支持向量机更差。处于实用的目的，选定支持向量机模型为最优模型进行当前数据集下的回归预测。

4. 论结论及进一步工作

本文通过分析对平台型电商转化率的影响因素的文献学习后，找到了两类重要且易得的数据作为自生变量，对转化率进行回归。比较不同的回归模型后发现，整体非线性模型的拟合效果更佳，其中支持向量机的拟合效果和稳定性最好，能够达到 91%的预测准确度。但需要申明的是不同机器学习模型在不同的数据量和特征维度下的性能是不同的，所以在一定时期后或者获得新的特征字段下，应该参考本文分析流程进行新的数据处理、特征选择、建模、对比和优化。

Table 1. Partial correlation table for order conversion rate and other variables
表 1. 订单转化率与其余变量的偏相关系数表

	UV	Turn	redu	off	Woff
Turn	-0.275	1	0.312	0.282	0.242
显著性(双侧)	0.001	0.001	0.001	0.001	0.001

Table 2. Model fitting accuracy index comparison table
表 2. 模型拟合准确度指标对比表

模型	训练集		测试集	
	平均误差平方和(MAE)	平均绝对误差率(MAPE)	平均误差平方和(MAE)	平均绝对误差率(MAPE)
多元线性回归	2.02097E-04	0.129307944	0.000288941	0.128367199
BP 神经网络	3.14678E-05	0.051443551	0.007998557	0.14762282
随机森林	3.86159E-05	0.048938157	0.000144018	0.090169743
支持向量机	7.33355E-05	0.070159606	0.000105093	0.076729838

本文存在的不足是在机器学习模型训练时，参数选择以 R 语言自带程序包中的函数 tune()函数给出的参数建议为准。在模型选择时，可能存在过度拟合的问题。为了得到更加精确地研究应当对应不同的回归模型，寻找到最佳的预测模型，再进行对比。本文在所具有的特征达到了不错的模型拟合效果，不过对于数据挖掘模型来说还存在数据量少和数据维度少的缺陷，或许在探索更多丰富的特征后可以得到更加优秀的预测效果，如线下广告投放、SKU 信息、促销商品信息、会员数等。另外可以尝试进行综合预测法，组合不同的预测方法以达到更好的预测效果。

参考文献

[1] 李双双, 陈毅文, 李江予. 消费者网上购物决策模型分析[J]. 心理科学进展, 2006, 14(2): 294-299.

[2] 刘贵容, 王哲, 林毅. 电商转化率影响因素分析与改进策略[J]. 商业时代, 2015(34): 72-74.

[3] 韩睿. 基于消费者感知的价格促销策略研究[D]: [博士学位论文]. 武汉: 华中科技大学, 2005.

[4] 李长春. 大数据背景下的商品需求预测与分仓规划[J]. 数学的实践与认识, 2017, 47(7): 70-79.

[5] 李永娜. 基于支持向量机的回归预测综述[J]. 信息通信, 2014(11): 32-33.

[6] 李静星. G 公司网上商城精准营销的研究[D]: [博士学位论文]. 广州: 广东财经大学, 2014.