基于机器学习 Xgboost 模型解决商店商品销量预测的问题

赵一安（湖北省武昌实验中学）

【中图分类号】F724.6 【文献标识码】A 【文章编号】1006-4222（2018）11-0250-03

1 引言

随着数据信息化的迅猛发展，数据量呈爆炸性增长。如何合理有效地利用企业掌握的数据为企业决策服务成为各企业关注的焦点。[1]因此，本文的研究课题主要针对销量预测。其既可以减少出现热销商品供不应求，冷门商品囤货堆积的现象，进而提高店铺利润。本文以 python 为主要工具进行数据分析，从多方面构建数据挖掘模型。基于已有的数据分析模型，以及我所掌握的实验所需要的商店商品近 5 年销量和各方面的数据，应用 Xgboost 模型挖掘不同维度的属性从而进行预测。

最终本文通过对数据的大规模分析，实现了对商品销量预测的模型的高完成度构建，进行了开创性的探究，便于店铺及时对进货量的调整，有利于减少通货膨胀和通货紧缩等状况出现的可能性，从而减少金融市场的波动，此外，更可减少货物滞销导致的资源浪费，对现代及未来商业的发展有着重要的意义。

2 国内外研究现状

计算机科学技术的高速发展的结果是计算机在各个行业都被有效的应用。而传统的销售、购物渠道也受到了极大的冲击。大数据时代到来使全球数据量级逐步增加，而互联网上信息的急剧增长，一方面使人们获取的信息来源愈来愈丰富，给人们生活带来极大的便利；另一方面，面对海量的信息资源，人们不得不花费更多的时间和精力去搜寻对其有用的信息，因此信息超载现象愈发严重，如何高效地利用数据也成为一个热点话题。

2.1 国外状况

针对国外状况，我们对美国、欧洲两个地区进行详细分析，从发展速度、规模、跨境电商、全渠道以及目前遇到的问题等几个方面来分析国外的情况。

2.1.1 美国线上、线下商店发展现状与趋势

据科尼尔管理咨询公司发布的《2015 年全球零售电子商务指数》显示，美国在 2015 年的全球零售电子商务市场排名中超过中国，以 2380 亿美元高居全球榜首。近几年来，美国电子商务发展速度有所放缓，但市场规模稳步扩大，社会零售总和交易额中的占比也随着每年的电商零售市场的交易额的提高呈现持续增长的态势。除此之外，线下实体店依然是消费者最青睐的购物渠道。

2.1.2 欧洲线上、线下商店发展现状与趋势

欧洲网络购物用户众多。电商市场所占比例大高速增长，按照西欧、北欧、中欧的顺序依次发展从高降低。作为全球最大的跨境电子商务市场区域的欧洲，其跨境电商发展相对成熟，市场规模已占到全球跨境电商市场规模的 35%，导致买家数量持续增长。然而，由于欧洲国家众多，不同的国家有其不同的法律体系，导致问题的发生。包括统一电子商务方面的立法，建立高度透明的统一支付标准；针对不同国家物流价格不透明的问题，提供统一的物流解决方案以及建立统一标准的增值税制度。

相较于线上店铺而言，欧洲线下店铺，包括传统奢侈品近年来销售额下滑，但一些历史更为悠久的欧洲奢侈品行业生产

教学中，教师可以采用实验原理学习、模拟操作、实物操作这一流程，增进学生对相关知识的理解，提升学生的实践操作能力，培养学生的逻辑思维。

3.2 学生在实验过程中可能偏离预定目标

虚拟实验室可以克服时间和空间的局限，学生能够运用互联网，获取大量信息资讯，有助于开阔学生的眼界。然而，过量的资料会增加学生筛选资料的难度，甚至可能出现偏离实验主题的情况。为此，教师要充分发挥引导职能，明确指出每次实验的主题目标，关于实验内包含的原理及知识点，应给予学生适当的指导，让学生明确重难点。

3.3 虚拟实验室对网络及硬件具有较高要求

学生要运用网络连接到虚拟实验平台，并完成网上作业，在长期使用之后，需要村村的实验数据会大幅度增多。如此一来，实验室对网络及配置的要求逐步提升。鉴于这一实际情况，相关管理人员要适当加大在维护和升级虚拟实验室上的资金，及时升级相应软件及硬件配置，并组织培养此方面的基础人才，科学管理相应基础设施，让实验资源得到科学配置和利用，充分发挥虚拟实验教学的作用。

4 总结

在高校人才培养规划中，如何培养学生的实践操作能力及创新品质历来是关键关节。当前，高校实验教学内容更新迅速，实验教学的重要性日愈凸显，但实验仪器设备却长期为得到更新、略显陈旧，且实验场地较为有限。这些问题均使得高校实验教学质量大打折扣，让培养学生实践能力及创新能力的目标难以实现。虚拟实验教学有效拓展了传统实验教学，凭借自身优势，让以上问题得到妥善解决。近年来，虚拟仿真技术日愈成熟，在高校实验室信息化建设进程中，虚拟实验室在其中扮演着不可忽视的作用。

参考文献

[1]彭永进，王昌军，赵晓艳，等.虚拟实验室在高等学校实验室信息化建设中的作用[J].中国管理信息化，2018(8).
[2]郑升华.虚拟化技术在高校实验室建设中的应用浅析[J].经营管理者，2017(14).

收稿日期：2018-10-13

仍属正常。为了提高销量,一些品牌、店铺开始进行更加大规模的生产。

2.2 国内状况

作为世界最大的电子商务市场，中国电商交易额近年来保持高速增长,已进入了一个平稳增长的发展阶段,同时电子商务企业也迎来了上市潮。随着中国城市化进程的继续,云计算、大数据、移动支付、信用体系的不断发展,物流行业正激发着整个社会创新的活力,电子商务行业也将继续受益。

通过数据分析了解客户购物行为的一些特征，对提高竞争力有极大帮助。利用数据挖掘技术通过对用户数据的分析,可以得到关于顾客购买取向和兴趣的信息，从而为商业决策提供可靠的依据。目前数据挖掘在市场销售上的应用可分为两类:数据库营销和货篮分析。数据库营销的任务是通过交互式查询、数据分割和模型预测等方法来选择潜在的顾客,以便向它们推销产品。数据库营销作为一种个性化的营销手段,可以有选择和有目的性地进行营销与客户关怀活动，从而扩大市场占有率与客户占有率,提升客户消费舒适度和满足感,进而增加顾客对企业品牌的忠诚度，取得企业与客户的双赢局面。货篮分析是利用市场销售数据来识别顾客的购买行为模式,从而帮助确定商店货架的布局摆放以促销某些商品,并且对进货的选择和搭配上也更有针对性。货篮分析常用于寻找顾客的购买模式,不同时期、不同地区的销售情况的对比等。与此同时，所收集的数据也非常有利于商品在未来的销售预测。相比之下,数据挖掘在商品销售预测方面的应用还有很大的发展空间。

3 数据来源与处理和模型假设

本研究选取了 Kaggle 网站(https://www.kaggle.com/)下载的九十万多条商品销量数据,这些数据囊括了十家商店,其中每家商店选取了 50 种商品,这些数据精确到五年内每种商品每日的销量情况。因此这份研究数据可以比较全面的反映这些商品近阶段的行情,确保了模型的广泛性与合理性,可以较为科学地预测其未来销量。

研究中所有的成绩数据和可能影响因素数据皆来自 Kaggle 网站。这些数据均存储在微软 Excel 表格中。

在建立模型之前,先对模型做出如下假设:

(1)假设每个商品的销量可以通过历史水平在一定程度上反映出来,即具有传承性,历史销量可以在一定程度上预测销量。

(2)假设商品销量在未来 3 个月所受到的影响因素与前五年不会有较大差距，即不会有紧急突发的状况导致的不可控因素的出现。

在此基础之上,使用 Xgboost 模型来进行分析。可将数据分为测试集和训练集,选取了九十万个数据来训练模型,测试集则选取了五万个数据,来检测模型的性能。

4 Xgboost 的算法介绍

众所周知,Xgboost 是“极端梯度上升”的简称,它类似于梯度上升框架,但是更加高效。来源于 Boosting 的它可谓是强有力的体现出了科学家对于强可学习、若可学习的研究。相较于它的“前辈”Adaboost,它采用了二阶泰勒展开以使算法可以更快收到最优。此外,它同时具有线性模型求解器和树学习算法的能力。因此,它快速的秘诀在于算法在单机上也可以并行计算的能力。这使得 Xgboost 至少比现有的梯度上升实现有至少 10 倍的提升。它提供多种目标函数,包括回归,分类和排序。

由于它在预测性能上的强大虽然相对缓慢的实现,“Xgboost”不失为一个理想选择。它还有做交叉验证和发现关键变量的额外功能。Xgboost 的思想是每次迭代学习之前学习后的损失值,用基分类器预测值相加,来预测结果。

提升算法的背后思想就如同:弱分类器就像偏科的同学,分类器学习的时候会选择该科目内应答能力最强的学生,最后就是一群偏科学生一起答完了一整张考卷。

值得提到的是，在近期多次机器学习竞赛中,Xgboost 可以说是最火的工具包之一,在靠前的排名中,它的“出场率”将近 70%。

5 使用 Xgboost 数据准备

(1)本次研究使用的特征包含以下内容:周几、几号、是否节假日、前 3d 平均、前 7d 平均、前 30d 平均、前 90d 平均、上周同天、上个月同天、3 个月前同天、6 个月同天(实际情况可能还需要更多的特征)。

(2)将所有其他形式的数据转换为数值型向量,再读取数据并观察其特征。

(3)数据清洗

为获得可训练用数据,我们对原始数据进行了清洗,具体过程如下:

①注意较为异常的数据。

②合并训练集和测试集，并添加新的用以区分训练集和测试集,值 1 为训练集,o 为测试集。

③数值化分类特征值。

④分解特征。

⑤规范化特征表达。

⑥删除残差大于 2.5 的异常数据。

⑦删除存在异常的数据点。

```
# 这里我想要筛选出的是开店后销售额却为零的异常点
df_all[df_all$sales=0|df_all$sales>0,'age']<--1
df_all$sales[df_all$sales=0]<-mean(df_all$age[df_all$age>0])
```

(4)特征处理

根据相关信息背景对数据进行特征处理:

Etc:添加一些额外的特征

6 模型的建立

(1)特征数据转换:将业务数据中取值可以简单枚举的字段转换成包含 1 个 1 其它均为 0 的 list (如：星期一为 1,0,0,0,0,0,0)(# 将数字转为 list)。

(2)使用特征数据建立模型

基本步骤:

```
# 读取特征和目标
# 获取训练、验证、测试的特征和目标(为了减少代码运行时间,方便测试)
# 取训练、验证、测试数据子集
# 训练、验证
# 解析训练、验证、测试数据的特征和目标
# 训练、验证数据转换为 Xgboost 矩阵
# 设置基本参数
# 创建训练模型
```

(根据迭代收敛情况,尝试着进行调参)

模型中,常用参数的详细说明如表 1。

(缺省值是当前系统可以获得的最大线程数)

默认数据即表 2。

7 实验结果及分析

集成方法:

本次实验使用的是较为简单的平均预测。

加权平均:

如果对预测值采用高惩罚措施,则可能造成过拟合,且本次试验是在相对“干净”的数据环境中进行的,因此我在最后

表 1

(1)booster:设置需要使用的上升模型。可选 gbtree(树)或 gblinear(线性函数),gbtree 是使用基于树的模型进行提升计算,gblinear 是使用线性模型进行提升计算,这里默认为 treebooster
(2)nthread:并行运行 xgboost 的线程数,输入的参数应该<系统的 CPU 核心数,若是没有设置算法会检测将其设置为 CPU 的全部核心数
(3)eta:收缩步长,即学习速率,可表示预计到达时间,取值范围是[0,1]。在更新叶子节点的时候权重乘以 eta,就可以在一定程度上避免在更新过程中会出现的过拟合现象。通常最后设置 eta 为 0.01~0.2eta 通过缩减特征的权重使提升计算过程更加保守。eta 值越大,则无法收敛的可能性越大,所以可以把 eta 的值设置的小些。
(4)max-depth:每棵树的最大深度,取值范围为[1,∞],缺省值为 6,就默认为 10,树越深,越容易过拟合。
(5)subsample:训练的实例样本占整体实例样本的比例,取值范围是(0,1],这里默认为 0.9。
(6)colsample-bytree:在构建每棵树时,列(特征)的子样本比,参数值的范围是(0,1]。
(7)objective:[default=reg:linear]定义学习任务及相应的学习目标,可选的目标函数如下:"reg:linear"-线性回归。"reg:logistic"-逻辑回归。"binary:logistic"-二分类的逻辑回归问题,输出为概率。"binary:logitraw"-二分类的逻辑回归问题,输出的结果为 wTx。"count:poisson"-计数问题的 poisson 回归,输出结果为 poisson 分布。
(8)seed:随机数种子,为确保数据的可重现性,可以默认为 0。
(9)①num_pbuffer[xgboost 自动设置,不需要设置]预测结果缓存大小,通常设置为训练实例的个数;②num_feature[xgboost 自动设置,不需要设置]在 boosting 中使用特征的维度。

表 2

参数	值	参数	值
objective	reg:linear	eta	0.015
booster	gbtree	max_depth	18
nrounds.cv	3000	subsample	0.7
nrounds.lb	270	colsample_bytree	0.7
eary.stop.round	100		

选取采用轻惩罚。

具体的模型如下:

线性模型 lm 拟合趋势,带特征交互的 glmnet 模型+Xgboost 模型拟合残差。

图 1 给出了 Xgboost 模型中非组合特征的重要度得分。

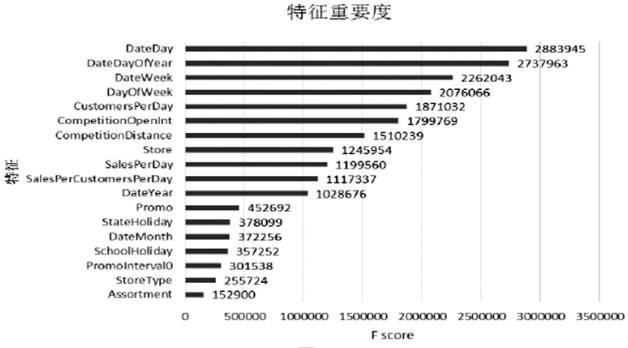

图 1

我们从图 2 中可以看出时间类特征子集和有关竞争对手的分值相较于其他项都非常的高,这意味着这些特征对模型具有非常大的影响,而商店种类差异与商品样式则可以不作为特征子集考虑。

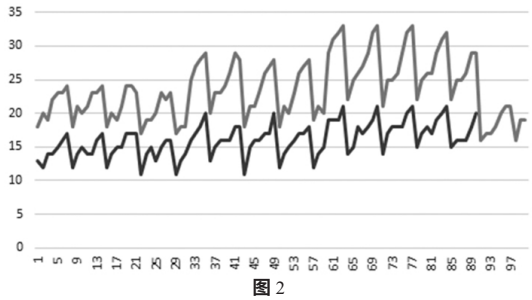

图 2

这是两种商品用 Xgboost 预测的未来三个月销量的折线图。图 2 描述了受调查的两种商品用 Xgboost 预测的未来三个月销量,可以清晰地看出,商品销量呈周期性增长,类似于生物学中的互利共生关系。可看出这两类商品的销量呈正相关。

图 3 为放入了多组商品数据的单 Xgboost 模型:

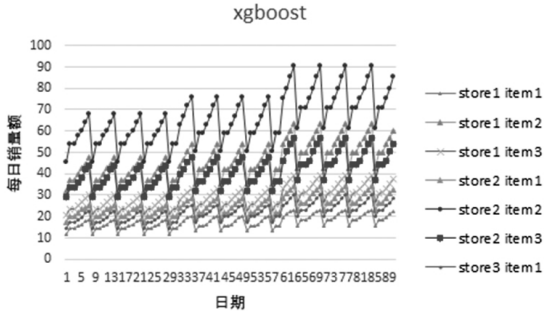

图 3

可以看出:

(1)大多数商店没有急剧向上或者向下的趋势。

(2)部分商店有两周的循环,受每隔一周进行的促销活动的影响。

(3)部分商店有季节性变化。

(4)大部分商店星期日和国定假日都是非营业日。

Xgboost 优缺点:

.Xgboost 在目标函数中显示的加上了正则化项。

.Xgboost 不仅使用到了一阶导数,还使用二阶导数。

.Xgboost 寻找分割点的标准是最大化。

…

总之,Xgboost 是一个高度复杂的算法,有足够的能力去学习数据的各种各样的不规则特征,相对于线性模型求解器和树学习算法,有过之而无不及。

8 结论

本论文研究建立于 Xgboost 方法对实体零售业销售额的预测的结果上。通过在特征工程中对原始数据进行特征提取、选择,筛选出所需要的对模型有一定影响的特征子集;本文通过特征处理与数据清洗,初步尝试了模型的集成学习方法和参数调优。这种基于单 Xgboost 的模型不仅适用于零售业销售额的预测,还可以拓展此方法以应用于国内零售实体业甚至电商平台的销售额预测,对于提高商店的运营生产模式、日常管理、价格管理、乃至精准营销具有一定的意义。[2]

参考文献

[1]赵啸彬.基于数据挖掘的零售业销售预测[D].上海:上海交通大学,2010.

[2]叶倩怡(导师:饶泓).基于 Xgboost 方法的实体零售业销售额预测研究[D].南昌大学,2016,05.

收稿日期:2018-10-17