中央财经大学

Central University of Finance and Economics

学年学期：__2020～2021 学年第一学期__

课程名称：_____数学与统计建模案例_____

课程代码：_____4012012_____

任课教师：_____杨欣欣_____

姓　　名：王云阁、吴宇翀、刘艳璐、宋智慧、杨莫依

班　　级：____数学与统计建模案例班____

组　　别：_____29 组_____

总　　分：_____

评 分 人：_____

摘要

我们通过案例研究分析某电商消费者购买行为数据，对消费者购买意向进行预测，通过建立多种算法模型，预测每个**消费者购买的概率**，并进行模型效果的比较。找到最具有普适性的预测模型，进而有效地推广到各种电商商品的销售，为各类型的销售者提供营销模式的参考依据。

我们在保持重抽样方法相同的情况下，使用多种常用机器学习模型对样本进行训练，与此同时，使用 10 折**交叉验证**进行模型比较。最终，**PLSDA** 模型在准确率和 Kappa 两项评判指标下都具有最好的效果。

在变量选择上，最重要的变量为：**网页价值**、**信息页访问时长**和**商品页访问时长**。对于所有的用户，这三个特征具有普适性，在预测用户是否购买中是较为重要的衡量因素。

关键词：**购买行为**，**分类预测模型**，**机器学习**，**变量选择**，**模型比较**

目录

1 背景

近年来，中国的电子商务快速发展，交易额连创新高，电子商务在各领域的应用不断扩展和深化、相关服务业蓬勃发展、支撑体系不断健全完善、创新的动力和能力不断增强。电子商务正与实体经济深度融合，进入规模性发展的阶段，对经济社会生活的影响不断增大，正成为我国经济发展的新引擎。整个社会的消费模式都因此产生了很大变化，从以实体店购物为主转变为足不出户的网络购物方式。2015年，中国电子商务市场交易规模达 16.4 万亿元，增长 22.7%。其中网络购物增长 36.2%，成为推动电子商务市场发展的重要力量。网络购物以其便利的操作方式、短时间等优势逐步成为居民购物的主要方式，目前仍维持着快速发展的趋势。线上购物凭借其庞大的客户群体且不断增长的购买方式，占 B2B（企业对企业）、B2C（企业对个人）和 C2C（个人对个人）市场收入的很大一部分。

据中商产业研究院整理，2019 年天猫 “双十一” 全天成交额为 2684 亿元，超 2018 年 549 亿元，再次创下新纪录。消费形式的转变是全球化趋势，据 2018 年 Optinmonster 公司的调查数据[1] 可知，有 69 % 的美国人每月都在网上购物，而 25 % 的美国人每月至少一次在网上购物。仅在美国，预计 2023 年将有 3 亿在线购物者，占全国人口的 91 %，且电子商务零售购买量预计将从 14.1 % 上升到 22 %。

在现今互联网大数据时代，随着电商的快速发展，分析用户购买意向数据对电商平台商品销量预测、确定商品营销范围，挖掘潜在用户等方面均有重要意义。影响用户购买意向的因素有很多，网页的访问数和访问时长、网页的访问时间和形式、线上购买时所用的操作系统和浏览器甚至也会对线上购物产生一定的作用。通过这些信息，企业和个人销售者可以了解和分析购物者线上购买特定商品的具体行为习惯以及各种外部因素以何种方式影响着商品销售。因此，销售者可以通过各影响因素之间的统计关系进行数据分析预测商品销量，合理安排市场营销方式，挖掘更多的潜在用户，进一步增加他们的销售和收入。

2 文献综述

不同学者通过不同的研究方法对用户的线上购买影响因素进行了分析。

袁和林等采用偏最小二乘方法通过建立 PLR-SEM 模型研究了顾客网购行为影响因素。通过分析发现个人属性的影响要强于电子服务因素。[1]

李宝库等通过回归分析、因子分析以及方差分析等方法研究了用户线上购买意向的影响因素。在用户网购行为的影响因素中，每个影响因素作用的大小不同，因此需要通过构建用户网购行为影响因素模型进而确定每个影响路径对应的系数。[2]

金灏利用分类与预测算法分析了用户的浏览行为和购买行为，实现了对潜在用户的挖掘。利用电商企业网站数据以及国内前期研究资料对本文所提出的数据挖掘处理计算方法进行实证模拟，研究企业如何实现对潜在客户相关信息的挖掘，促进潜在客户转变为企业的现实客户、忠实客户。

[1] https://optinmonster.com/online-shopping-statistics/

3 研究方法

1. 首先对数据进行分析和处理，建立模型对消费者购买意向进行预测。其次，选用多种机器学习模型进行模型比较，为研究影响消费者购买意向的因素提供更多的评判思路。找到最具有普适性的预测模型，进而有效地推广到各种电商商品的销售，为各类型的销售者提供营销模式的参考依据。
2. 我们通过应用多种机器学习模型如 Logit 回归、线性判别分析（LDA）、偏最小二乘判别分析（PLSDA）、SVM、随机梯度助推法等方法，探究影响消费者购买意向的因素，将重要的变量筛选出来，理清楚其影响关系，使用数据集中的这些变量预测商品销售。并且将数据向更深的层次进行挖掘，探究内在的关系。
3. 通过分析各变量之间的关系，找出变量之间是否有相关性，提高模型的准确性。最后通过数据可视化的方式，利用各种图表将变量之间存在的联系直观的展现出来。
4. 该研究通过分析消费者购买意向的影响因素为商品销售的预测提供思路，可以根据实际情况加以调整。

4 数据集与描述分析

4.1 数据集说明

我们使用一个公开的数据集[2] ，它有 35 个变量，310 个观测。[3]

数据集由分属于 12330 个会话的特殊向量组成。在数据集的 12330 个会话中，其中 84.5%（10422）是以购物结束的负类样本，其余（1908）是以购物结束的正类样本。数据集的形成使得每个会话在一年的时间内属于不同的用户，以避免出现特定活动、特殊日期、用户配置文件或时段的趋势。

数据集由 10 个数值属性和 8 个分类属性组成。“Revenue” 特征可用作类标签。“Administrative”、“AdministrativeDuration“、“Informational”、“Informational Duration”、“Product Related” 和 “Product Related Duration” 特征代表访问者在该会话中访问的不同类型页面的数量以及在这些不同类别的页面中花费的总时间。这些特征的值来源于用户访问的页面的 URL 信息，并在用户进行实际操作时实时更新，例如从一个页面移到另一个页面。“Bounce Rate”、“Exit Rate”、“Page Value” 代表了 “谷歌分析” 对电子商务网站中每个页面的度量。Bounce Rate 的值是在该会话期间从该页面进入站点然后离开（“跳出”）而不触发对分析服务器的任何其他请求的访问者的百分比。特定网页的 Exit Rate 的值是该页的所有页面浏览量，即会话中最后一个页面的百分比。Page Value 是用户在完成电子商务交易之前访问的网页的平均值。Special Day 是指网站访问时间接近某个特定的日子（如母亲节、情人节），在这一天，会议更有可能最终完成交易。此属性的值是通过考虑电子商务的动态（如订单日期和交货日期之间的持续时间）来确定的。例如，对于情人节，该值在 2 月 2 日和 2 月 12 日之间取一个非零值，在此日期之前和之后为零，除非它接近另一个特殊的日期，否则它的最大值出现在 2 月 8 日值为 1。数据

[2]数据来源: http://archive.ics.uci.edu/ml/datasets/Online+Shoppers+Purchasing+Intention+Dataset#

[3]模型的变量取值和分布见附录

集还包括操作系统、浏览器、区域、流量类型、访客类型（返回或新访客）、布尔值，指示访问日期是周末还是一年中的月份。

表 1: 变量解释和类型

变量名	变量描述	数据格式
Administrative	主页访问数	Number
Administrative Duration	主页访问时长	Number
Informational	信息页访问数	Number
Informational Duration	信息页访问时长	Number
ProductRelated	商品页访问数	Number
ProductRelated Duration	商品页访问时长	Number
BounceRates	跳出率，指通过该页面进入网站并在不触发任何其他任务的情况下退出网站的访问者百分比。	Number
ExitRates	退出率，在该特定页面结束网站浏览量百分比。	Number
PageValues	网页价值，网页价值是用户在登陆目标页面或完成电子商务交易（或两者）之前访问的网页的平均值。	Number
SpecialDay	特殊日，此值表示浏览日期与更可能完成交易的特殊日期或假日（例如母亲节或情人节）的接近程度。	Number
Month	月份，浏览网页的月份。	String
OperatingSystems	操作系统，表示用户在查看页面时所在的操作系统。	Integer
Browser	浏览器，表示用户用来查看页面的浏览器。	Integer
Region	地区，表示用户位于哪个区域。	Integer
TrafficType	流量类型，表示用户归类为哪种类型的流量。	Integer
VisitorType	访问者类型，表示访问者是“新访问者”, “回访者”还是“其他”。	String
Weekend。	周末，表示用户操作是否在周末	Factor
Revenue	购买，表示用户是否完成购买。	Factor

4.2 数据预处理

1. 对原始数据进行去重补缺等预处理。
2. 我们将用户被转化**购买**，与**未购买**相对应，生成一个虚拟变量。

4.3 描述分析

4.3.1 跳出率/退出率

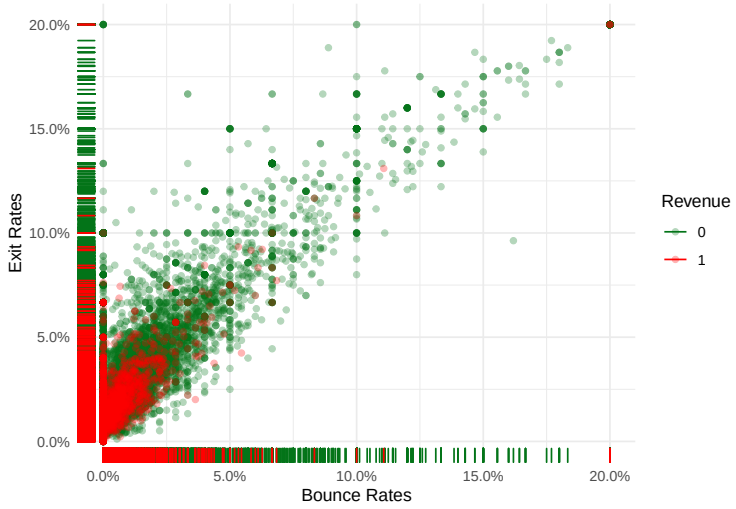

图 1: 未购买与购买两类用户跳出率、退出率分布密度图（红色代表购买）

购买的用户跳出率、退出率均较低，集中在 0-5% 之间。未购买的用户的跳出率、退出率分布较分散，大部分集中在 0-10% 之间。跳出率、退出率是是否购买的一个较为重要的衡量指标。

4.3.2 周末、网页价值与购买行为

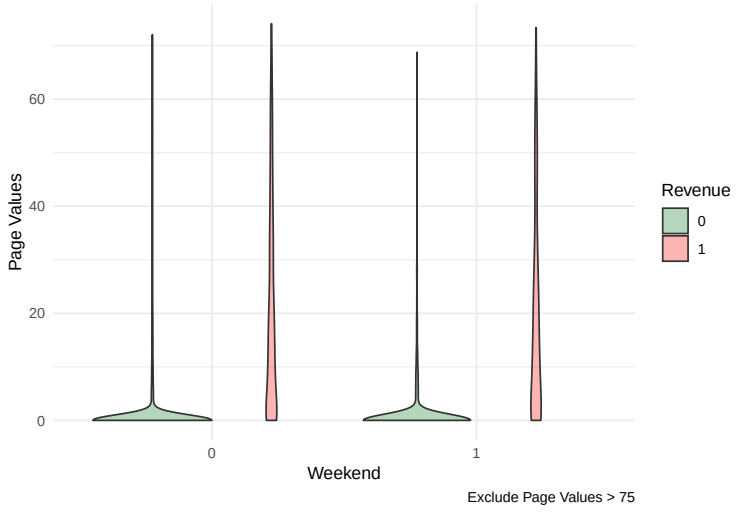

图 2: 周末网页价值分布密度图（红色代表购买）

由图可知，无论用户最后是否购买，用户是否在周末操作网页价值差别不大。我们认为周末这一特征对用户的购买影响较小，网页价值与用户购买的相关性也较低。

4.3.3 临近特殊日商品页访问与购买行为

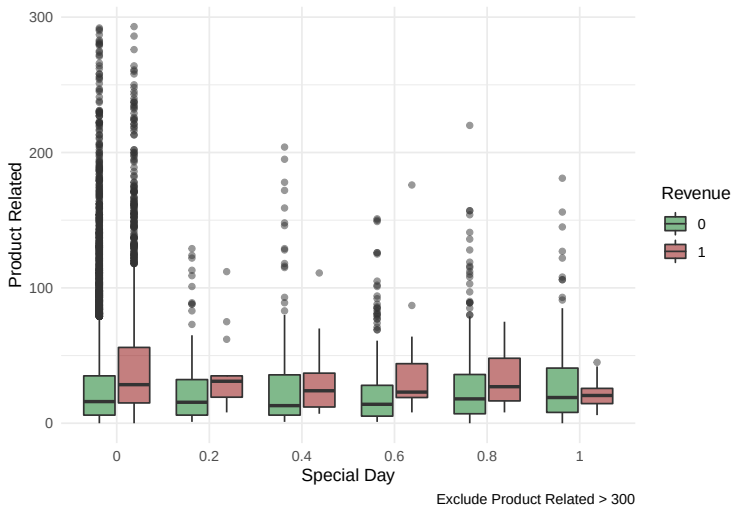

图 3: 购买与不购买两类用户临近特殊日商品页访问数箱线图（红色代表购买）

首先，购买的用户商品页访问数高于不购买的用户。用户访问更多的商品页代表着用户有更强烈的购买需求或购买欲望。越接近特殊日期，无论用户最后最终是否购买，离群值均显著增加，说明用户越接近特殊日期，访问的商品页会显著增多。因此，特殊日期是促进用户购买的一个重要因素。

4.3.4 临近特殊日的用户类型分布

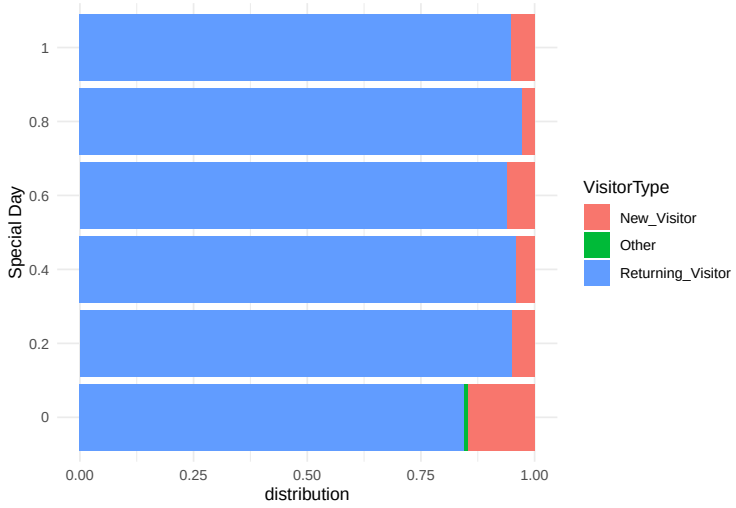

图 4: 临近特殊日的用户类型分布（红色代表购买）

由图可知：

1. 回访者的贡献度要远远高于其他两种访问者。
2. 接近特殊日期，新用户的贡献度增加。
3. 其他类型的访问者贡献度几乎为 0。

可见是否是回访者是衡量用户是否购买的重要因素之一。

5 建立解释模型

5.1 拟合

我们将**购买**作为响应变量，建立 logit 回归模型。选取的自变量有：

主页访问数 / 主页访问时长 / 信息页访问数 / 信息页访问时长 / 商品页访问数 / 商品页访问时长 / 跳出率 / 退出率 / 网页价值 / 特殊日 / 月份

我们将因子型变量转换成隐变量后加入模型中，连续型变量直接加入模型。

根据 Z 检验的 p 值可知[4] ，“商品页访问时长”、“退出率”、“网页价值” 以及月份中的 “二月”、“三月”、“五月”、“十一月” 和 “十二月” 在统计上显著。[5]

退出率高的用户购买的可能性较低。退出率是指，对于某一特定页面而言，从该页面离开网站的访问占所有浏览到该页面的访问的百分比。用户从该页面退出的比率越高，可能意味着用户对该页面商品的兴趣度越低，因此购买的概率也随之降低。

5.2 预测

我们划分四分之三的训练集和四分之一的验证集。

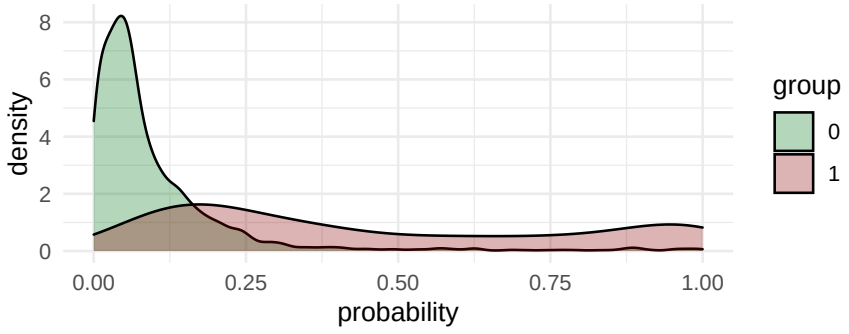

图 5: 预测的购买概率值（红色代表购买）

[4]模型详细见附录

[5]在 95% 置信区间下

从预测概率分布图可知，对于未购买的用户，我们预测出的购买概率值比较低，但对于购买的用户，预测出的购买概率值比较分散，超过一半的预测购买率低于 50%。

我们猜想：**模型讲未购买的用户预测为购买的用户的概率比较地，因此比较难识别出购买的用户。**

5.3 混淆矩阵与验证结果

我们将预测概率大于 50% 的判定为购买。

灵敏度（Sensitivity）

$$\text{灵敏度} = \frac{\text{正确判定为“购买”的样本数量}}{\text{观测到的“购买”的样本数量}}$$

特异度（Specificity）

$$\text{特异度} = \frac{\text{正确判定为“购买”的样本数量}}{\text{观测到的“购买”的样本数量}}$$

假购买率

$$\text{假购买率} = 1 - \text{观测到的“购买”的样本数量}$$

表 2: 混淆矩阵表

Prediction	Reference	Freq
0	0	2534
1	0	71
0	1	286
1	1	191

表 3: 验证结果表

Accuracy	Kappa	AccuracyLower	AccuracyUpper	AccuracyNull	AccuracyPValue	McnemarPValue
0.884	0.457	0.872	0.895	0.845	0	0

表 4: 灵敏度和特异度等指标表

	Sensitivity	Specificity	Pos Pred Value	Neg Pred Value	Precision
指标值	0.4	0.973	0.729	0.899	0.729

使用 Logit 回归模型进行预测的准确率大致为: 88.4% ，准确率较高。

由灵敏度可知，40% 的有购买倾向的顾客会被模型成功捕捉到；由特异度可知，模型的误判率只有 2.7%。模型可以捕捉到购买的顾客，同时模型预测认为会购买的顾客有极大的概率会进行购买。

5.4 接受者操作特征（ROC）曲线

我们使用 ROC 曲线 (Altman 和 Bland 1994; Brown 和 Davis 2006; Fawcett 2006) 决定分类概率的阈值。[3] [4] [5]

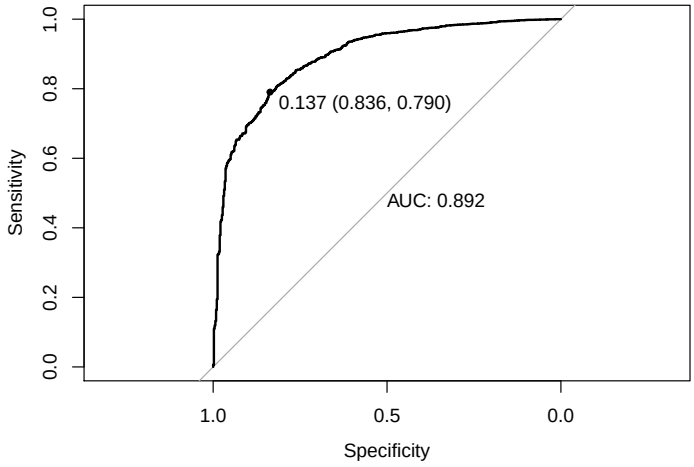

图 6: Logit 模型的 ROC 曲线

通过降低阈值可以达到提高灵敏度的目的，但同时也承担着特异度降低的风险，导致误判率上升。在实际操作中，可以通过使用不同阈值的方法达到不同的效果:

1. 当业务要求尽可能减低误判率时，则可以选择适当提高阈值以达到目的。
2. 当业务要求尽可能识别出会购买的用户时，则可以选择适当降低阈值以达到目的。

6 预测模型的选择

6.1 抽样、训练与评价指标

我们使用 Kappa 统计量（Cohen 1960）作为模型准确度的度量指标。[6]

$$\text{Kappa} = \frac{O - E}{1 - E}$$

上述公式中，O 代表准确性，E 则代表的是根据混淆矩阵边缘计数得出的期望准确性。1 值表示模型的预测与观测类是相同的，0 值意味着观测类和预测类是不同的，该统计量取值是在-1 和 1 之间，其中负数代表实际和预测值是相反的，但实际情况中，绝对值较大的负数值在模型的预测中出现的频率非常低。在各类分布相同的时，总精确度与 Kappa 成比例。Kappa 值在 0.30 到 0.50 之间，代表一致性合理，但这一取值区间也要依具体情况而定。（Agresti 2002）

6.2 Logit 回归

表 5: 在重抽样下 Logit 模型的表现

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.889	0.472	0.022	0.121

Logit 的应用非常广泛，因为该模型非常简单，并且计算速度很快，而且具有很强的可解释性。尽管逻辑回归模型的预测分类能力较好，但如果我们仅着重于预测准确性这一衡量指标，可以找到表现更好的模型。

6.3 线性判别分析（LDA）

我们使用 Fisher（1936）[7] 和 Welch（1939）[8] 提出的最优判别准则的方式。

通过贝叶斯法则，已知：

$$\Pr\left[Y = C_\ell | X\right] = \frac{\Pr\left[Y = C_\ell\right] \Pr\left[X | Y = C_\ell\right]}{\sum_{\ell=1}^{C} \Pr\left[Y = C_\ell\right] \Pr\left[X | Y = C_\ell\right]}$$

若：

$$\Pr\left[Y = C_1\right] \Pr\left[X | Y = C_1\right] > \Pr\left[Y = C_2\right] \Pr\left[X | Y = C_2\right]$$

将 X 分入 C_1，得到线性判别函数为：

$$X'\Sigma^{-1}\mu_\ell - 0.5\mu_\ell'\Sigma^{-1}\mu_\ell + \log\left(\Pr\left[Y = C_\ell\right]\right)$$

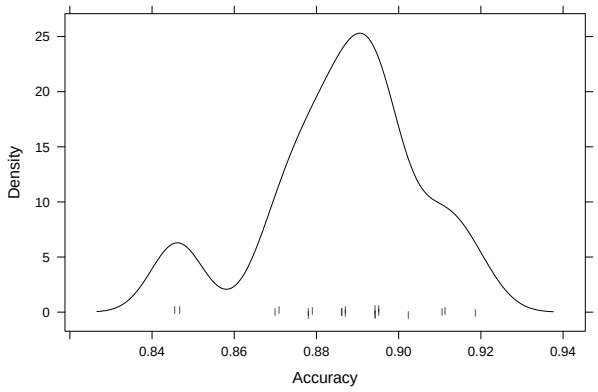

图 7: 在重抽样下 LDA 模型的准确率分布

6.4 偏最小二乘判别分析（PLSDA）

当变量之间有较强的多重共线性，LDA 模型便不再适用。我们尝试通过使用主成分分析来压缩变量空间的维度。这一方法的缺点是，PCA 可能无法识别能将样本分类的较好的变量组合，同时，由于 PCA 是无监督学习，我们很难通过它找到一个最优的分类预测。

Berntsson 和 Wold（1986）[9] 提出了偏最小二乘判别分析（PLSDA）。尽管 Liu 和 Rayens（2007）[10] 指出，在不降维的情况下，LDA 一定优于 PLS。但降维后 PLS 的表现可能超过 LDA。

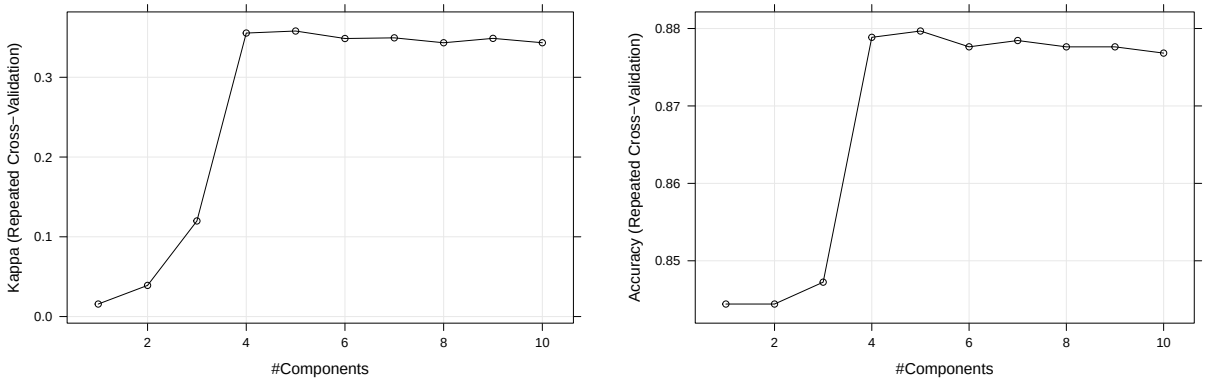

图 8: Kappa 指标和准确率随主成分个数的变化

由图可知，随主成分个数的增多，Kappa 指标先上升，之后稍有下降；随着主成分个数的增加，准确率先下降，后上升到顶峰、再下降。在此模型中，对于 Kappa 指标和准确率指标，选取前 4 个主成分都是最优的。

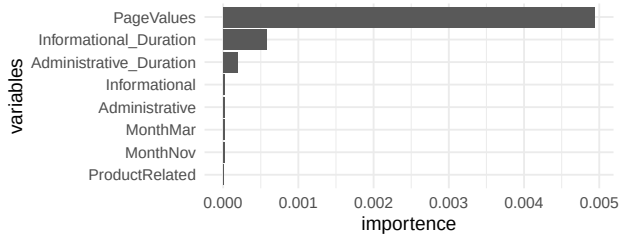

图 9: 变量重要程度

变量以在 PLSDA 模型中的重要性为标准进行排序，重要度排名前三位的分别是：Pagevalue 网页价值，Informational Duration 信息页访问时长，ProductRelated Duration 商品页访问时长。**对于所有的用户，这三个特征具有普适性，在预测用户是否购买中是较为重要的衡量因素**。

而重要程度最低的三个变量分别是 MonthJul、MonthDec、MonthJune。这三个变量对用户购买商品没有太大的影响。

6.5 SVM

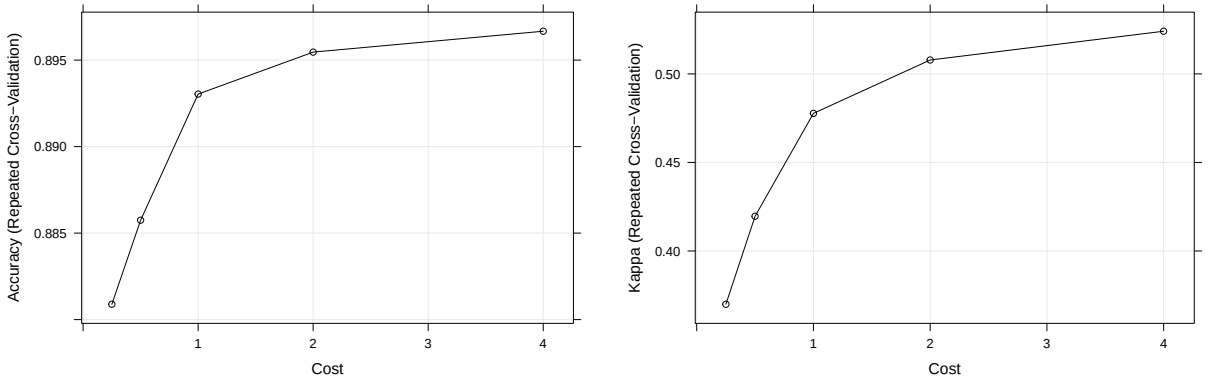

图 10: 调优参数不同取值下的准确率和 Kappa 指标变化

在损失参数增大的同时，准确率指标与 Kappa 指标的变化趋势相同，准确率和 Kappa 值均呈现上升趋势。

6.6 随机梯度助推法（GBM）

我们使用 Friedman 等（2000）[11] 提出的通过最小化指数损失函数实现分类的方式，构建随机梯度助推模型。

$$f_i^{(0)} = \log \frac{\hat{p}}{1-\hat{p}}$$

其中，$f(x)$ 为预测值，$\hat{p}_i = \frac{1}{1+\exp[-f(x)]}$

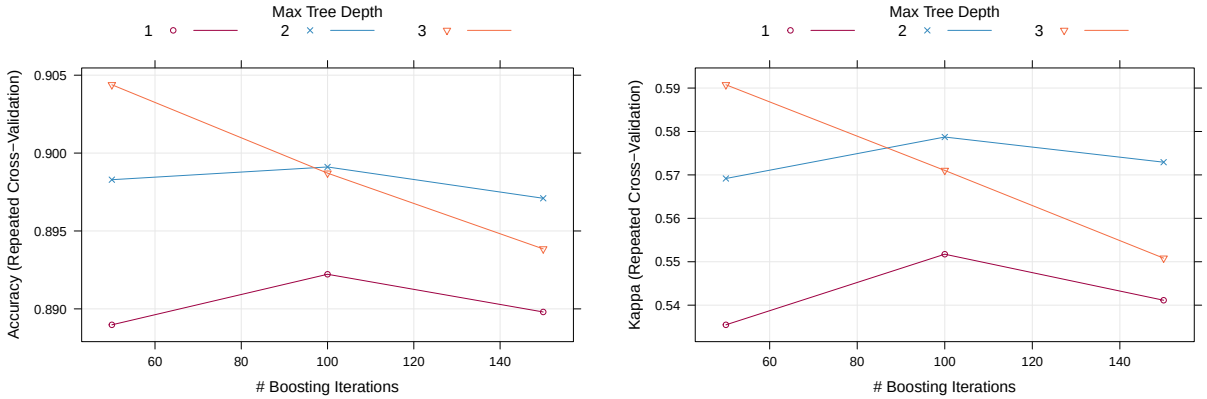

图 11: 调优参数和迭代次数不同取值下的准确率和 Kappa 指标变化

当迭代次数为 1 次和 2 次时，随着助推树的加深，Kappa 值和准确率均呈现先上升后下降的趋势。迭代两次的 Kappa 值和准确率高于迭代一次的 Kappa 值和准确率。但迭代次数为 3 次时，随着树的加深，Kappa 值和准确率呈下降趋势。

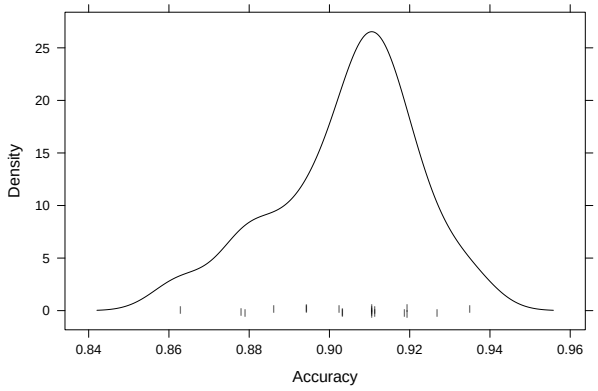

图 12: 在重抽样下 GBM 模型的准确率分布

6.7 模型间的比较

所有模型都使用相同的重抽样方法，且我们保证不同模型使用的重抽样样本完全一致。

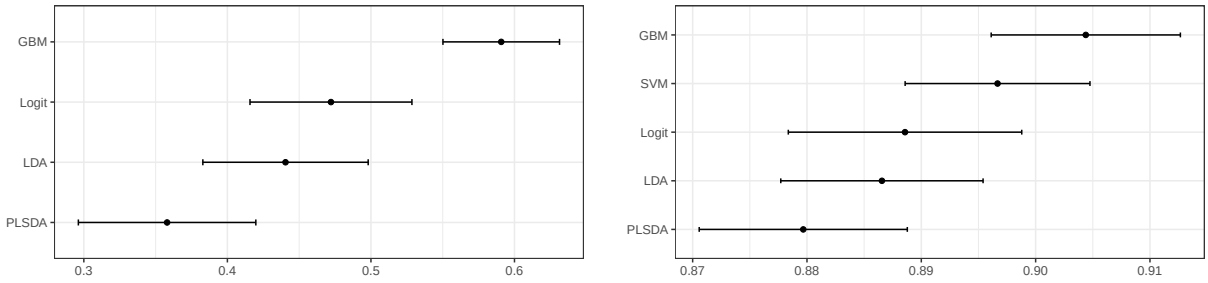

图 13: 模型间准确率和 Kappa 的比较（0.95 置信区间）

若由 **Kappa** 值来衡量，GBM 模型效果最好，Logit 和 LDA 模型次之，PLSDA 模型效果最差。

若由**准确率**来衡量，从偏差的角度来看，GBM 模型效果最好，SVM 模型次之；从方差的角度来看，GBM 模型、SVM 模型方差较小，PLSDA 模型和 logit 模型方差较大。

7 结论

在此研究中，我们主要研究了电商用户的购买行为。我们通过这个研究一个案例，所得出的相关结论，对电商平台的营销决策提供一定的建议。

于此同时，此研究建立的多种预测购买模型，完全可以在电商的其它领域中适当地调整后加以应用。

7.1 变量解释

1. 购买的用户跳出率、退出率均较低，集中在 0-5% 之间。未购买的用户的跳出率、退出率分布较分散，大部分集中在 0-10% 之间。跳出率、退出率是是否购买的一个较为重要的衡量指标。
2. 无论用户最后是否购买，用户是否在周末操作网页价值差别不大。我们认为周末这一特征对用户的购买影响较小，网页价值与用户购买的相关性也较低。
3. 购买的用户商品页访问数高于不购买的用户。用户访问更多的商品页代表着用户有更强烈的购买需求或购买欲望。越接近特殊日期，无论用户最后最终是否购买，离群值均显著增加，说明用户越接近特殊日期，访问的商品页会显著增多。因此，特殊日期是促进用户购买的一个重要因素。
4. 回访者的贡献度要远远高于其他两种访问者。接近特殊日期，新用户的贡献度增加。其他类型的访问者贡献度几乎为 0。可见是否是回访者是衡量用户是否购买的重要因素之一。

7.2 模型选择

若由 **Kappa 值**来衡量，GBM 模型效果最好，Logit 和 LDA 模型次之，PLSDA 模型效果最差。

若由**准确率**来衡量，从偏差的角度来看，GBM 模型效果最好，SVM 模型次之；从方差的角度来看，GBM 模型、SVM 模型方差较小，PLSDA 模型和 logit 模型方差较大。

综合来看，**GBM 模型**具有最好的效果。

7.3 变量选择

在 PLSDA 模型中的各变量重要性排序：排在前三名的是 Pagevalue 网页价值，Informational Duration 信息页访问时长，ProductRelated Duration 商品页访问时长。它们是在预测用户是否购买中是较为重要的衡量因素。而月份变量重要性较低，对用户购买商品没有太大的影响。

8 参考文献

[1] YUAN D, LIN Z, ZHUO R. What drives consumer knowledge sharing in online travel communities?: Personal attributes or e-service factors?[J]. Computers in Human Behavior, 2016, 63: 68–74.

[2] 李宝库, 刘莹. 农村居民网络消费溢价支付意愿研究 [J]. 中国流通经济, 2019, 33(2): 103–112.

[3] ALTMAN, DOUGLAS, G., 等. Diagnostic tests 3: receiver operating characteristic plots.[J]. Bmj British Medical Journal, 1994.

[4] BROWN C D, DAVIS H T. Receiver operating characteristics curves and related decision measures: A tutorial[J]. Chemometrics & Intelligent Laboratory Systems, 2006, 80(1): 24–38.

[5] FAWCETT T. An introduction to ROC analysis[J]. Pattern Recognition Letters, 2006, 27(8): 861–874.

[6] COHEN J A. A Coefficient of Agreement for Nominal Scales[J]. Educational & Psychological Measurement, 1960, 20(1): 37–46.

[7] FISHER R A. The Use of Multiple Measurements in Taxonomic Problems[J]. Annals of Eugenics, 1936, 7(7): 179–188.

[8] L. W B. (ii) Note on Discriminant Functions[J]. Biometrika, 1939(1-2): 1–2.

[9] BERNTSSON P, WOLD S. Comparison Between X-Ray Crystallographic Data and Physicochemical Parameters with Respect to Their Information about the Calcium Channel Antagonist Activity of 4-Phenyl-1,4-dihydropyridines[J]. Quantitative Structure Activity Relationships, 1986, 5(2): 45–50.

[10] LIU Y, RAYENS W. PLS and dimension reduction for classification[J]. Computational Statistics, 2007, 22(2): 189–208.

[11] BEN-DOR, AMIR, BRUHN, 等. Tissue Classification with Gene Expression Profiles[J]. Journal of Computational Biology, 2000.

[12] 吴林武. 电子商务个性化推荐系统对消费者购买意向的影响研究 [D]. 天津大学, 2019.

[13] 代倩宇. 移动社交电商用户购买意愿影响因素研究 [D]. 华中师范大学, 2019.

[14] 苏秦, 李钊, 崔艳武, 等. 网络消费者行为影响因素分析及实证研究 [J]. 系统工程, 2007, 25(2): 1–6.

9 附录

9.1 模型间准确率和 Kappa 的比较

表 6: 模型间准确率的比较

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LDA	0.846	0.878	0.887	0.887	0.895	0.919	0
PLSDA	0.839	0.868	0.879	0.880	0.886	0.919	0
SVM	0.862	0.893	0.899	0.897	0.903	0.927	0
GBM	0.863	0.894	0.911	0.904	0.913	0.935	0
Logit	0.839	0.884	0.891	0.889	0.903	0.919	0

表 7: 模型间准确率差异矩阵

	LDA	PLSDA	SVM	GBM	Logit
LDA		0.006881	-0.010113	-0.017830	-0.002019
PLSDA	0.406844		-0.016994	-0.024712	-0.008900
SVM	0.367854	0.026244		-0.007717	0.008094
GBM	0.001186	2.984e-05	0.638866		0.015811
Logit	1.000000	0.435876	0.860707	0.003634	

表 8: 模型间 Kappa 的比较

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.	NA's
LDA	0.179	0.352	0.442	0.440	0.530	0.635	0
PLSDA	0.162	0.244	0.364	0.358	0.416	0.603	0
SVM	0.266	0.491	0.535	0.524	0.576	0.676	0
GBM	0.388	0.545	0.610	0.591	0.645	0.751	0
Logit	0.245	0.407	0.498	0.472	0.571	0.635	0

表 9: 模型间 Kappa 差异矩阵

	LDA	PLSDA	SVM	GBM	Logit
LDA		0.006881	-0.010113	-0.017830	-0.002019
PLSDA	0.406844		-0.016994	-0.024712	-0.008900
SVM	0.367854	0.026244		-0.007717	0.008094
GBM	0.001186	2.984e-05	0.638866		0.015811
Logit	1.000000	0.435876	0.860707	0.003634	

9.2 模型训练指标详情

表 10: 在重抽样下 LDA 模型的表现

parameter	Accuracy	Kappa	AccuracySD	KappaSD
none	0.887	0.44	0.019	0.123

表 11: 在重抽样下 PLSDA 模型的表现

ncomp	Accuracy	Kappa	AccuracySD	KappaSD
1	0.844	0.016	0.005	0.036
2	0.844	0.039	0.008	0.056
3	0.847	0.120	0.019	0.118
4	0.879	0.355	0.019	0.130
5	0.880	0.358	0.019	0.132
6	0.878	0.349	0.017	0.120
7	0.878	0.349	0.017	0.121
8	0.878	0.343	0.017	0.123
9	0.878	0.349	0.017	0.121
10	0.877	0.343	0.016	0.112

表 12: 在重抽样下 SVM 模型的表现

sigma	C	Accuracy	Kappa	AccuracySD	KappaSD
0.061	0.25	0.881	0.370	0.017	0.125
0.061	0.50	0.886	0.420	0.016	0.103
0.061	1.00	0.893	0.478	0.018	0.117
0.061	2.00	0.895	0.508	0.018	0.113
0.061	4.00	0.897	0.524	0.017	0.101

表 13: 在重抽样下 GBM 模型的表现

	shrinkage	interaction.depth	n.minobsinnode	n.trees	Accuracy	Kappa	AccuracySD	KappaSD
1	0.1	1	10	50	0.889	0.535	0.024	0.104
4	0.1	2	10	50	0.898	0.569	0.019	0.102
7	0.1	3	10	50	0.904	0.591	0.018	0.087
2	0.1	1	10	100	0.892	0.552	0.023	0.104
5	0.1	2	10	100	0.899	0.579	0.019	0.095
8	0.1	3	10	100	0.899	0.571	0.023	0.110
3	0.1	1	10	150	0.890	0.541	0.024	0.111
6	0.1	2	10	150	0.897	0.573	0.019	0.087
9	0.1	3	10	150	0.894	0.551	0.021	0.094

9.3 Logit 回归结果

表 14: Logit 回归系数表

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.88	0.17	-11.25	0.00
Administrative	0.00	0.01	0.35	0.73
Administrative_Duration	0.00	0.00	-0.56	0.58
Informational	0.03	0.03	1.10	0.27
Informational_Duration	0.00	0.00	0.37	0.71
ProductRelated	0.00	0.00	1.13	0.26
ProductRelated_Duration	0.00	0.00	2.23	0.03
BounceRates	-4.53	3.34	-1.36	0.17
ExitRates	-16.86	2.38	-7.08	0.00
PageValues	0.08	0.00	34.13	0.00
SpecialDay	-0.13	0.24	-0.56	0.58
MonthDec	-0.60	0.18	-3.32	0.00
MonthFeb	-1.82	0.64	-2.85	0.00
MonthJul	0.08	0.22	0.36	0.72
MonthJune	-0.32	0.27	-1.17	0.24
MonthMar	-0.52	0.18	-2.89	0.00
MonthMay	-0.57	0.17	-3.30	0.00
MonthNov	0.54	0.16	3.32	0.00
MonthOct	0.01	0.20	0.04	0.97
MonthSep	0.01	0.21	0.03	0.98

```
##
## Call:
## glm(formula = Revenue ~ Administrative + Administrative_Duration +
##     Informational + Informational_Duration + ProductRelated +
##     ProductRelated_Duration + BounceRates + ExitRates + PageValues +
##     SpecialDay + Month, family = binomial(link = "logit"), data = dat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -6.1120  -0.4669  -0.3408  -0.1629   3.5081
##
## Coefficients:
```

```
##                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)              -1.882e+00  1.673e-01 -11.251  < 2e-16 ***
## Administrative            3.835e-03  1.095e-02   0.350 0.726141
## Administrative_Duration -1.084e-04  1.936e-04  -0.560 0.575700
## Informational             2.968e-02  2.690e-02   1.103 0.269967
## Informational_Duration    8.065e-05  2.208e-04   0.365 0.714905
## ProductRelated            1.285e-03  1.141e-03   1.127 0.259846
## ProductRelated_Duration   5.977e-05  2.680e-05   2.230 0.025726 *
## BounceRates              -4.533e+00  3.341e+00  -1.357 0.174871
## ExitRates                -1.686e+01  2.380e+00  -7.082 1.42e-12 ***
## PageValues                8.198e-02  2.402e-03  34.134  < 2e-16 ***
## SpecialDay               -1.323e-01  2.369e-01  -0.558 0.576707
## MonthDec                 -6.016e-01  1.814e-01  -3.316 0.000914 ***
## MonthFeb                 -1.819e+00  6.385e-01  -2.849 0.004393 **
## MonthJul                  7.833e-02  2.180e-01   0.359 0.719360
## MonthJune                -3.214e-01  2.741e-01  -1.173 0.240938
## MonthMar                 -5.196e-01  1.795e-01  -2.894 0.003802 **
## MonthMay                 -5.725e-01  1.733e-01  -3.303 0.000957 ***
## MonthNov                  5.372e-01  1.620e-01   3.316 0.000914 ***
## MonthOct                  7.194e-03  2.013e-01   0.036 0.971489
## MonthSep                  6.113e-03  2.123e-01   0.029 0.977024
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 10624.8  on 12329  degrees of freedom
## Residual deviance:  7179.6  on 12310  degrees of freedom
## AIC: 7219.6
##
## Number of Fisher Scoring iterations: 7
```

9.4 数据指标明细

```
## 'data.frame':    12330 obs. of  18 variables:
##  $ Administrative         : int  0 0 0 0 0 0 0 1 0 0 ...
##  $ Administrative_Duration: num  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational          : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Informational_Duration : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ ProductRelated         : int  1 2 1 2 10 19 1 0 2 3 ...
##  $ ProductRelated_Duration: num  0 64 0 2.67 627.5 ...
##  $ BounceRates            : num  0.2 0 0.2 0.05 0.02 ...
##  $ ExitRates              : num  0.2 0.1 0.2 0.14 0.05 ...
##  $ PageValues             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ SpecialDay             : num  0 0 0 0 0 0 0.4 0 0.8 0.4 ...
##  $ Month                  : Factor w/ 10 levels "Aug","Dec","Feb",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ OperatingSystems       : Factor w/ 8 levels "1","2","3","4",..: 1 2 4 3 3 2 2 1 2 2 ...
##  $ Browser                : Factor w/ 13 levels "1","2","3","4",..: 1 2 1 2 3 2 4 2 2 4 ...
##  $ Region                 : Factor w/ 9 levels "1","2","3","4",..: 1 1 9 2 1 1 3 1 2 1 ...
##  $ TrafficType            : Factor w/ 20 levels "1","2","3","4",..: 1 2 3 4 4 3 3 5 3 2 ...
##  $ VisitorType            : Factor w/ 3 levels "New_Visitor",..: 3 3 3 3 3 3 3 3 3 3 ...
##  $ Weekend                : Factor w/ 2 levels "0","1": 1 1 1 1 2 1 1 2 1 1 ...
##  $ Revenue                : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...

##  Administrative   Administrative_Duration Informational
##  Min.   : 0.000   Min.   :   0.00         Min.   : 0.0000
##  1st Qu.: 0.000   1st Qu.:   0.00         1st Qu.: 0.0000
##  Median : 1.000   Median :   7.50         Median : 0.0000
##  Mean   : 2.315   Mean   :  80.82         Mean   : 0.5036
##  3rd Qu.: 4.000   3rd Qu.:  93.26         3rd Qu.: 0.0000
##  Max.   :27.000   Max.   :3398.75         Max.   :24.0000
##
##  Informational_Duration ProductRelated   ProductRelated_Duration
##  Min.   :   0.00        Min.   :  0.00   Min.   :    0.0
##  1st Qu.:   0.00        1st Qu.:  7.00   1st Qu.:  184.1
##  Median :   0.00        Median : 18.00   Median :  598.9
##  Mean   :  34.47        Mean   : 31.73   Mean   : 1194.8
##  3rd Qu.:   0.00        3rd Qu.: 38.00   3rd Qu.: 1464.2
##  Max.   :2549.38        Max.   :705.00   Max.   :63973.5
##
##   BounceRates         ExitRates         PageValues        SpecialDay
##  Min.   :0.000000   Min.   :0.00000   Min.   :  0.000   Min.   :0.00000
##  1st Qu.:0.000000   1st Qu.:0.01429   1st Qu.:  0.000   1st Qu.:0.00000
##  Median :0.003112   Median :0.02516   Median :  0.000   Median :0.00000
##  Mean   :0.022191   Mean   :0.04307   Mean   :  5.889   Mean   :0.06143
##  3rd Qu.:0.016813   3rd Qu.:0.05000   3rd Qu.:  0.000   3rd Qu.:0.00000
##  Max.   :0.200000   Max.   :0.20000   Max.   :361.764   Max.   :1.00000
##
##      Month      OperatingSystems    Browser         Region      TrafficType
```

```
##  May    :3364   2      :6601     2      :7961   1      :4780   2      :3913
##  Nov    :2998   1      :2585     1      :2462   3      :2403   1      :2451
##  Mar    :1907   3      :2555     4      : 736   4      :1182   3      :2052
##  Dec    :1727   4      : 478     5      : 467   2      :1136   4      :1069
##  Oct    : 549   8      :  79     6      : 174   6      : 805   13     : 738
##  Sep    : 448   6      :  19     10     : 163   7      : 761   10     : 450
##  (Other):1337   (Other):  13     (Other): 367   (Other):1263   (Other):1657
##             VisitorType    Weekend  Revenue
##  New_Visitor      : 1694   0:9462   0:10422
##  Other            :   85   1:2868   1: 1908
##  Returning_Visitor:10551
##
##
##
##
```