Operations Research and Fuzziology 运筹与模糊学, 2023, 13(5), 5650-5658
Published Online October 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.135564

Hans汉斯

线上旅游产品销量的影响因素分析和预测

滕斯琦

南京信息工程大学数学与统计学院，江苏 南京

收稿日期：2023年9月5日；录用日期：2023年10月17日；发布日期：2023年10月26日

摘　要

随着信息技术的飞速发展，在线旅游成为一种新兴的商业模式，在线旅游网站具有非常丰富的产品信息，包括产品价格、消费者评论、旅行社介绍等，其中每一项指标都会对线上旅游产品的销量产生不同程度的影响。本论文使用八爪鱼采集器对携程旅行网的在线产品的相关信息进行收集。使用Logit定序回归的方法，筛选出影响历史销量的12个关键因素，包括目的地类型、旅游类型、产品评分、点评数量、产品价格、旅行社评分、精致小团、赠取消险、无购物、攻略完备、立即确认和限时促销。最后，分别使用决策树模型和随机森林模型对产品销量进行了预测，参数调优后随机森林模型的AUC为0.9749。相较于决策树模型(AUC = 0.9190)，随机森林模型显示出更高的预测精度，能够更好地区分正例和负例样本，具备更出色的分类性能。

关键词

定序回归，随机森林模型，决策树模型，在线旅游产品

Analysis and Forecasting of Influencing Factors of Online Travel Product Sales

Siqi Teng

School of Mathematics and Statistics, Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Sep. 5th, 2023; accepted: Oct. 17th, 2023; published: Oct. 26th, 2023

Abstract

With the rapid development of information technology, online tourism has become an emerging business model. Online travel websites have a wealth of product information, including product prices, consumer reviews, travel agency introductions, and more. Each of these indicators will have varying degrees of impact on the sales of online travel products. This thesis utilizes the Octopus

文章引用: 滕斯琦. 线上旅游产品销量的影响因素分析和预测[J]. 运筹与模糊学, 2023, 13(5): 5650-5658.
DOI: 10.12677/orf.2023.135564

Collector to gather relevant information about the online products from Ctrip.com, a renowned travel website. By using Logit fixed-order regression, 12 key factors that affect historical sales were filtered out. These factors include destination type, tour type, product rating, number of reviews, product price, travel agent rating, exquisite small group, complimentary cancellation insurance, no shopping, complete strategy, immediate confirmation, and limited-time promotion. Finally, the decision tree model and random forest model were utilized to predict product sales. After parameter tuning, the random forest model achieved an AUC of 0.9749. When compared to the decision tree model (AUC = 0.9190), the random forest model demonstrates higher prediction accuracy, better differentiation between positive and negative samples, and superior classification performance.

Keywords

Ordinal Regression, Random Forest Model, Decision Tree Model, Online Travel Products

1. 引言

在线旅游(OTA，全称 Online Travel Agency)是一种在互联网平台上进行的以旅游产品购买和旅游服务预订为主要内容的消费模式。OTA 的各个平台都会推出一些具有特色的旅游产品和活动，并具有消费群体年轻化、旅游消费质量化的特征，这些特征可以为消费者提供个性化的服务，并对他们的出游和目的地决策产生积极的影响。多数消费者已形成了经常浏览在线旅行社平台的习惯，较高的用户粘性将会带来较多的旅游商机。当前，携程、同程、去哪儿网三家公司占据了国内在线旅游市场的主导地位。在激烈的竞争中，各个在线旅游平台都将注意力集中在了用户流量和市场份额上，各个旅行社通过提高服务质量、丰富产品等方式，推出满足消费者各种不同需求的产品[1]。

近几年来，由于我们的消费理念发生了改变，线上旅游消费逐渐占据了大部分人的生活方式，对 OTA 产品的研究日益受到国内外学者的重视。国外学者更多地从网站和用户两个视角，对网站的建设、历史销量、以及购买决策这几个方面展开了深入的研究。我国在旅游电子商务领域的研究起步较晚，更多的是从用户接受行为等方面来探讨，很少从技术运用的角度来探讨旅游电子商务。在研究方法方面，国内学者使用的大多是经验研究，其中大多为描述性研究。国外对电子商务的研究，更多的是结合多种处理方法，并辅以过程跟踪、文本挖掘等技术来进行数据收集，从而使得结论更有说服力[2]。

有关 OTA 的研究，多数集中于顾客评论对其交易量的影响，包括评论的内容、数量等[3]，相对单一化；此外，现有研究多关注于影响因素，缺乏对在线旅游产品建立预测模型。所以，本文对在线网站(携程旅游网)的数据进行了挖掘，采集从南京出发到国内华东、华南、华中、华北、西南、东北和港澳台这七个目的地的不同路线的各项数据信息，使用真实的数据，选取对 OTA 产品历史销量有影响的因素，构建更加合理、更加有效的算法模型，从而可以更加准确地对线上旅游产品的销售进行预测。更深层次地归纳出了旅游消费者的某些偏好，同时为在线旅游网站与旅行社提出了有效的建议[4]。

2. 数据来源及预处理

2.1. 数据来源

本文在携程网站中利用八爪鱼采集器爬取了以南京作为出发地，时间为 2023 年 6 月，旅游目的地分

别为华东、华南、华中、华北、西南、东北和港澳台 7 个大区的在线旅游产品的详细数据信息，其中数据信息包括销量、旅行社评分、旅程天数、旅程类型、产品评分、点评数量、产品售价、目的地、产品优惠、服务承诺这 10 个变量。从携程网站中共爬取了 4375 个在线旅游产品的数据，删除重复数据以及异常数据后得到容量为 3099 的有效样本。将上述除销量外的 9 个自变量归为如下 5 个类别。

(1) 价格信息：即各产品的标出价格。

(2) 用户反馈：包括旅行社的评分、产品的评分和点评数量三个方面。

(3) 行程特色信息：包括旅程的天数、旅程的类型和目的地类型三个方面。

(4) 优惠活动信息：包括暑期特惠、暑期早鸟特惠、精致小团和限时促销。

(5) 服务保障信息：有无购物、无自费、成团保障、攻略完备、亲子甄选、自选酒店、免费接送、提前 2 天免费退、立即确认和赠取消险十类。

2.2. 数据预处理

由于携程网站上的一些旅游产品条目缺少一些参数性信息，或在数据爬取时出现错误，导致原始数据中出现了一些需要处理的无用数据。原始数据中有 287 条信息缺失 10 个参数中的 3 个以上，由于各旅行产品随机且独立，且就样本总量而言比例较小，故直接删去整条记录。另外有 989 条重复数据，所以也删去整条记录，留下 3099 条有效信息。对于旅程天数、产品评分、点评数量、产品价格的缺失数据，使用均值法进行插补填充。最后，对数据进行了一致性检验，通过对原始资料的分析，并未发现显著的异常值，因此最终得到 3099 个有效样本。

2.3. 数据变换

数据变换是指将数据从一种表达形式转变为另一种的过程，主要目标是把数据转换成便于分析的形态。针对不同的分析模式，在数据类型上有不同的要求。本论文以 OTA 旅游商品的历史销售数据为因变量，其他 9 个变量为自变量，为了提高模型的精度，首先对其中的部分变量进行转换。

(1) 产品价格：根据第一四分位数(2060.5 元)和第三四分位数(4614.5 元)，将产品售价划分为低、中、高三种价位，以便进行描述性统计分析。

(2) 旅程天数：根据第一四分位数(4 天)和第三四分位数(6 天)，将旅程在 4 天以下的设为短期，将 4 至 6 天之间的设为中期，将 6 天以上设为长期。

(3) 旅行社评分：根据第一四分位数(4.7 分)和第三四分位数(4.9 分)，将旅行社划分为三个档次，其中 4.7 分以下的为一般，4.7 分~4.8 分的为普通，4.9 分及以上的为优秀。

(4) 产品评分：为了确保该变量的可行性和便于分析，将产品评分划分为 5 分、4.6 分~4.9 分、4.5 分及以下这三个档次。其中，产品评分主要集中在 5 分，共 1342 条数据，占比 43.3%，其次为 4.9 分，共有 758 条数据。

(5) 点评数量：按照第一四分位数(3 条)和第三四分位数(52 条)，将点评数量划分为少、中等、多三种。

(6) 产品优惠：产品优惠以字符串形式呈现，共包括四个水平，将该产品所含有的产品优惠水平用“1”表示，未含的则标注为“0”。

(7) 服务承诺：服务承诺以字符串形式呈现，共包括十个水平，将该产品所含有的服务承诺水平用“1”表示，未含的则标注为“0”。

3. 研究方法

3.1. 定序回归

回归分析是一种用来分析因变量 Y 与自变量 X 是否具有相关关系以及分析相关程度的常用方法。而

定序回归是一种回归分析方法，其因变量Y为定序变量，自变量X则是一般的解释变量[5]。

当模型中假定随机误差服从logistic分布时定序回归又称Logit定序回归。假设定序变量$Y(Y=1,2,...,J)$由J个类别构成，下面给出Logit定序回归方程的表达式：

$$\begin{aligned} L_j(X) &= logit\left[F_j(X)\right], j=1,2,...,J-1 \\ &= \log\left\{P(Y\le j\mid X)/\left[1-P(Y\le j\mid X)\right]\right\} \end{aligned} \tag{1}$$

其中，$F_j(X)=P(Y\le j\mid X)$为J类别的累积概率函数。

若Y独立于X，有：

$$L_j(X)=\alpha_j \tag{2}$$

其中，α_j为常数，若Y与X不独立，有：

$$L_j(X)=\alpha_j+\beta X \tag{3}$$

其中，β为系数。当$\beta>0$时，$L_j(X)=\alpha_j+\beta X$代表在固定数值X的条件下，在高次序一端出现的累计概率函数随的X增长而增加，反之亦然。

3.2. 随机森林

随机森林(Random Forest) [6]是一种基于集成学习的分类与回归算法。它以决策树为基本单位，在每个子集上构造多个决策树，通过bootstrap重采样产生不同的样本子集。在构建决策树时，随机森林从全部特征中随机选择一部分特征进行节点分裂。最后，通过采用众数表决法或取平均法来得出最终的预测结果。在随机森林的构建过程中，首先使用行抽样法对样本进行抽样，这可能导致有重复抽样的情况出现。然后进行列采样，通过使用完全分裂采样数据构造决策树，以确保叶节点无法再进一步分裂。这种随机性的引入有助于减少过拟合的风险，并提高模型的泛化能力。

4. 数据建模

4.1. 线上旅游产品历史销量的定序回归模型

构建Logit定序回归模型，对关键变量进行提取。首先将历史销量以四分位数为节点转化成有序变量，并将各分类变量数值化，具体的变量赋值情况如表1。

Table 1. Ordered regression model variable assignments
表1. 定序回归模型变量赋值情况

变量	赋值
历史销量	1：高销量 2：中销量 3：低销量
目的地类型	1：华东 2：华南 3：华中 4：华北 5：西南 6：东北 7：港澳台
旅行天数	1：长期 2：中期 3：短期
旅游类型	1：半自助游 2：跟团游 3：私家团 4：游学 5：自由行
产品评分	1：高 2：中 3：低
点评数量	1：多 2：中 3：少
产品价格	1：高 2：中 3：低
旅行社评分	1：优秀 2：普通 3：一般

对于经过处理的有序变量数据集，我们建立了一个Logit定序回归模型，并对该模型进行广义似然比检验，表2中列出了相应的检验结果。从表2中可以看出，P值(Pr)为0，这意味着整个模型的显著性非常高，即在Logit定序回归下，至少有一个自变量对因变量(销量)有显著的影响。

Table 2. Logit sequencing regression generalized likelihood ratio test results
表 2. Logit 定序回归广义似然比检验结果

Model	Resid. df	Resid. Dev	Test	Df	LR stat.	Pr (Chi)
1	3096	6413.579				
2	3075	3003.116	1vs2	21	3410.463	0

接下来，对每个自变量的显著性展开讨论。在 Logit 定序回归下，进行每个自变量显著性的假设检验，使用 R 语言构建定序回归全模型，并在表 3 中呈现定序回归全模型的单因素方差分析结果。

Table 3. One-way ANOVA results of the whole model of ordered regression
表 3. 定序回归全模型单因素方差分析结果

	LR Chisq	Df	Pr (>Chisq)	Signif
目的地类型	0.95	1	0.330176	
旅行天数	21.86	1	2.932e−06	***
旅游类型	25.98	1	3.449e−07	***
产品评分	2649.67	1	< 2.2e−16	***
点评数量	51.27	1	8.034e−13	***
产品价格	4.20	1	0.040525	*
旅行社评分	1.91	1	0.166564	
暑期特惠	0.26	1	0.609313	
暑期早鸟特惠	1.74	1	0.187536	
精致小团	28.91	1	7.573e−08	***
限时促销	5.38	1	0.020393	*
无购物	2.77	1	0.095763	
无自费	0.64	1	0.424173	
成团保障	8.09	1	0.004451	**
攻略完备	0.03	1	0.871297	
亲子甄选	0.11	1	0.745173	
自选酒店	2.08	1	0.148756	
免费接送	0.21	1	0.648094	
提前 2 天免费退	4.44	1	0.035135	*
立即确认	2.28	1	0.131466	

结果说明在 0.05 显著水平下，旅行天数、旅游类型、产品评分、点评数量、产品价格、精致小团、限时促销、成团保障、提前 2 天免费退这 9 个变量是显著的。由以上研究发现，所选择的变量的确可以对网上旅行产品的销售做出某种解释。但总体来看，整个模型的构建过于复杂，去除不显著的变量，可以使模型得到简化和改善。采用逐步回归的方法，依据 AIC 信息统计量，建立 Logit 定序回归模型，表 4 展示了模型结果。

Table 4. Results of Logit ordered regression AIC model
表 4. Logit 定序回归 AIC 模型结果表

	Value Std	Error	t value
目的地类型	−0.1015	0.01930	−5.257
旅游类型	−0.1651	0.03377	−4.890
产品评分	−0.1999	0.04328	−4.617
点评数量	2.1862	0.05006	43.675
产品价格	−0.3687	0.04095	−9.004
旅行社评分	0.1352	0.04397	3.074

Continued

精致小团	−0.1182	0.06878	−1.719
限时促销	−0.6298	0.10844	−5.808
无购物	0.3104	0.08009	3.876
攻略完备	−0.4528	0.14068	−3.219
立即确认	0.9291	0.58003	1.602
赠取消险	−0.1117	0.05784	−1.932

从上表可以看出，定序回归 AIC 模型中的自变量包括目的地类型、旅游类型、产品评分、点评数量、产品价格、旅行社评分、精致小团、限时促销、无购物、攻略完备、立即确认、赠取消险这 12 个变量，这些显著变量即为影响产品销量的关键因素，也将用于后续的建模分析当中。

进一步对 Logit 定序回归模型所筛选出的 12 个关键变量展开相关性分析，并绘制出相关热力图，如图 1 所示。

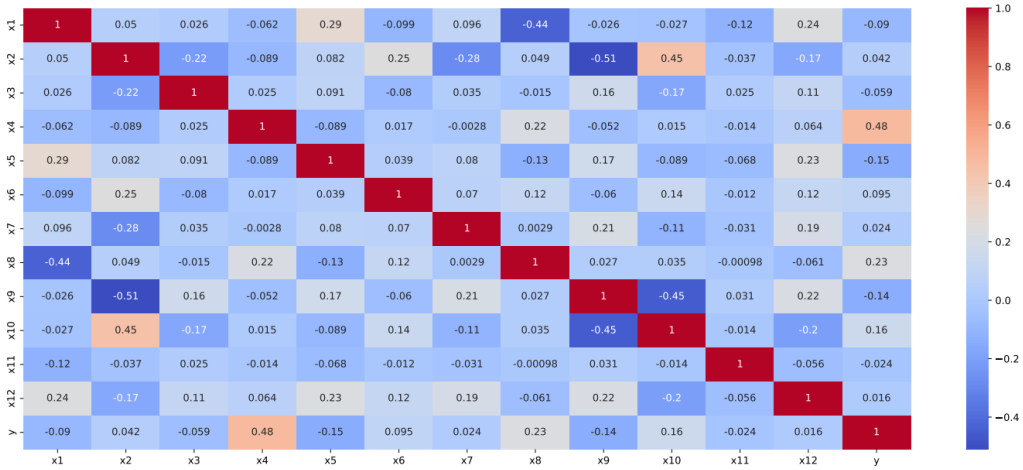

Figure 1. A heat diagram of the correlation between variables
图 1. 各变量之间的相关关系热能图

4.2. 在线旅游产品销量的随机森林预测模型

本文使用定序回归筛选出的关键变量，建立随机森林模型作进一步的分析和预测，表 5 为 RF 模型的变量说明。

Table 5. Description of random forest model variables
表 5. 随机森林模型变量说明

变量名	变量说明
历史销量(Y)	以均值为节点将销量分为低销量(0)和高销量(1)
目的地类型(X1)	华东、华南、华中、华北、西南、东北、港澳台
旅游类型(X2)	半自助游、跟团游、私家团、游学、自由行
产品评分(X3)	在 2~5 之间
点评数量(X4)	在 1~11357 之间
产品价格(X5)	在 516~24893 之间
旅行社评分(X6)	在 2~5 之间
精致小团(X7)	包含取值为 1，不包含取值为 0
限时促销(X8)	包含取值为 1，不包含取值为 0

Continued

无购物(X9)	包含取值为 1，不包含取值为 0
攻略完备(X10)	包含取值为 1，不包含取值为 0
立即确认(X11)	包含取值为 1，不包含取值为 0
赠取消险(X12)	包含取值为 1，不包含取值为 0

整个操作步骤如下：

(1) 在 Pycharm 编辑器中导入处理后的数据导入，并存储为 Dataframe 格式，最后以 7:3 的比例划分训练集和测试集，其中训练集包含 1859 数据，测试集包含 1240 条数据。

(2) 在训练集上，使用 Random Forest Regression 函数训练随机森林模型，随后在测试集上进行测试，最后对参数进行调整，得到最优参数的随机森林模型。

(3) 采用十折交叉验证法，验证模型结果，并计算出随机森林模型的准确率。

表 6 为使用网格搜索法之后得到的随机森林模型的各个参数。

Table 6. Random forest parameter settings table
表 6. 随机森林参数设置表

criterion	max_depth	max_features	min_samples_split	n_estimators
gini	5	0.4	12	18

采用调参后构建的随机森林模型对销量进行预测，并计算精确率、召回率、准确率及 f1-score 来衡量模型的性能、评估不同类别的预测结果。具体结果如下图 2 所示。

随机森林精确度...

	precision	recall	f1-score	support
0	1.00	0.98	0.99	1101
1	0.87	0.98	0.92	139
accuracy			0.98	1240
macro avg	0.93	0.98	0.95	1240
weighted avg	0.98	0.98	0.98	1240

Figure 2. Random forest prediction results
图 2. 随机森林预测结果

使用 sklearn.metrics 中的 confusion matrix 函数绘制随机森林分类器的混淆矩阵，结果如图 3 所示。

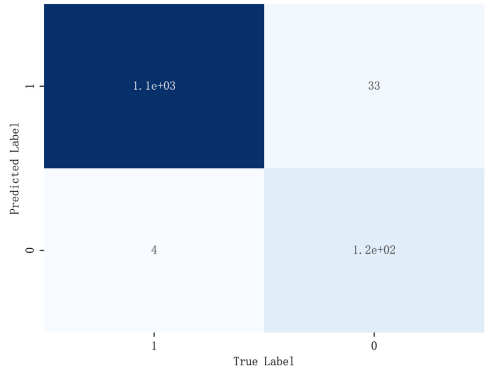

Figure 3. Random forest model confusion matrix
图 3. 随机森林模型混淆矩阵

图中轴上的“0”“1”分别代表“低销量”和“高销量”，根据混淆矩阵图可以得到，随机森林模型对于实际为低销量样本存在部分预测失误，对于实际为低销量且预测为低销量的样本和实际为高销量且预测为高销量的样本有更高的预测精确度，误分类样本数较少。

4.3. 与其他分类器(决策树)进行对比

ROC (Receiver Operating Characteristic)曲线是用于评估二分类模型性能的一种常用工具。它展示了二分类模型在不同阈值下的真阳性率(True Positive Rate，也称为灵敏度)和假阳性率(False Positive Rate)之间的关系。ROC 曲线越靠近左上角，说明模型的性能越好。除了 ROC 曲线，我们还可以使用曲线下的面积(Area Under the Curve, AUC)值来衡量模型的性能，AUC 的取值范围在 0 到 1 之间，它越接近 1，表示模型的预测准确性越高。

将上述数据集在决策树分类器上实现，采用 AUC 评价指标对决策树模型和随机森林模型的分类效果进行对比。其中决策树模型的 ROC 曲线如图 4 左图所示，随机森林模型的 ROC 曲线如图 4 右图所示。

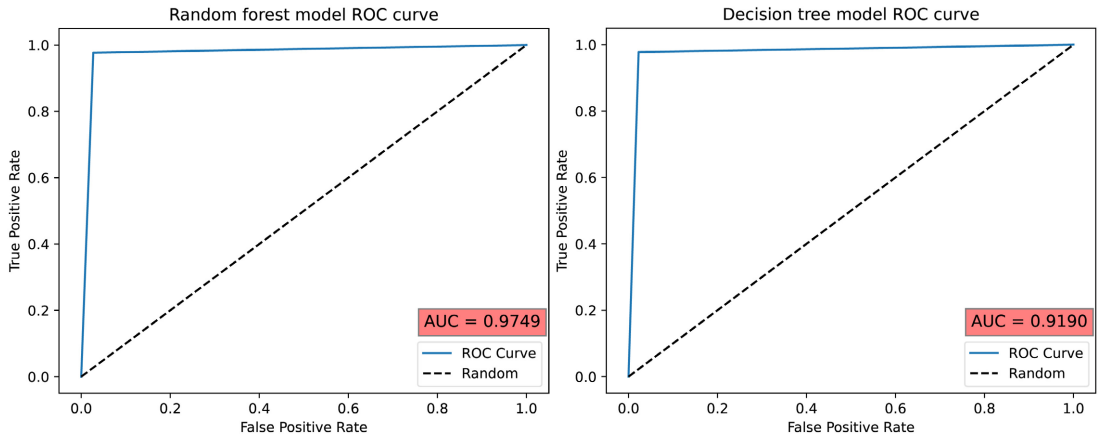

Figure 4. Decision Tree Model ROC Curve (Left) Random Forest Model ROC Curve (Right)
图 4. 决策树模型 ROC 曲线(左)随机森林模型 ROC 曲线(右)

从图 4 中的 AUC 值来看，随机森林模型的 AUC 为 0.9749，接近于 1。相较于决策树模型(AUC = 0.9190)，随机森林模型显示出更高的预测精度，能够更好地区分正例和负例样本，具备更出色的分类性能。

5. 结论与讨论

本文使用八爪鱼采集器对携程旅行网的在线旅游产品的相关信息进行收集，利用用 Logit 定序回归的方法，筛选出影响产品历史销量的 12 个关键因素。结合筛选出的关键因素，分别使用决策树模型和随机森林模型对产品销量进行预测，经参数调优后，随机森林模型的 AUC 为 0.9749，相较于决策树模型(AUC = 0.9190)，预测精度更高，能够更好地区分正例和负例样本，具备更出色的分类性能。

随机森林模型虽然在本文中展现了出色的预测性能，但仍存在一些不足之处。首先，随机森林模型在处理高维稀疏数据时可能会面临困难，因为其基本单位决策树在这种情况下往往无法发挥较好的表现。其次，由于随机森林是一种集成学习方法，其中包含多个决策树，因此模型的建立和训练过程相对较慢，对大规模数据集的处理可能存在一定的挑战。未来研究可以进一步探索如何提升随机森林模型在高维稀疏数据下的性能，例如引入特征选择方法或对决策树进行优化。此外，可以考虑结合其他机器学习技术或深度学习模型，以构建更强大的集成模型，从而进一步提升预测精度。

参考文献

[1] 司首婧. 线上旅游产品购买决策影响因素分析——基于 DEMATEL 方法[J]. 重庆科技学院学报(社会科学版), 2019(4): 57-59.

[2] 魏宝祥, 王亚林. 近十年旅游电子商务研究综述[J]. 江苏商论, 2017(8): 43-50.

[3] 夏美玉. 在线评论对在线旅游产品销量的影响研究[D]: [硕士学位论文]. 南昌: 江西财经大学, 2019.

[4] 方晓莹. 旅游平台产品销量的影响因素分析和预测[D]: [硕士学位论文]. 大连: 东北财经大学, 2022.

[5] 杨蕊岚. 基于定序回归的云终端用户行为可信性评估方法的研究[D]: [硕士学位论文]. 北京: 北京工业大学, 2017.

[6] 郑婕. 基于随机森林和 XGBoost 算法的二手车价格预测[J]. 数字技术与应用, 2021, 39(6): 90-93.