基于随机森林和相似度匹配算法的用户学习产品消费行为价值分析

摘 要：在互联网高速发展背景下，商业社会与互联网紧密联系，线上渠道与线下渠道相结合进行经营已经成为了各行业的共识。这需要利用数据挖掘技术对用户的行为数据进行分析，构建模型用于判别用户的价值，对不同用户制定专门的营销策略，从而以较低的营销成本实现用户转化率的提升。

对数据首先进行预处理，提高数据质量。获取的数据共包括四张表，即用户信息表、用户登录情况表、用户访问统计表和用户下单表。首先对获取的数据利用 python 软件进行数据的规约。数据的预处理过程包括处理存在缺失值、指标数值异常、指标范围错误等异常样本以提升数据的质量。接着对用户的各城市分布情况和登录情况进行分析，并分别将结果进行可视化。本文利用**区域地图、三维五角类地图以及三维热力图**对于所有用户的城市分布进行可视化展现，并通过 SPSS 软件分析得出一线城市下单率更高的结论。用户登录情况利用 EXCEL 和 SPSS 软件，采用整体分析同局部分析相结合的方法进行分析。整体分析中对用户登录情况进行频数分析，局部分析中对已下单和暂未下单用户的登录情况进行对比分析。构建模型判断用户最终是否会下单购买或下单购买的概率。本文利用预处理后的数据划分 70%训练样本集和 30%测试样本集，构建随机森林模型，给出各特征重要性，并利用网格搜索优化模型参数，选择在交叉验证下的最高平均准确度为 98.61%的模型作为最终模型，同时泛化能力达到 98.64%。构建用户行为相似度匹配模型，将中心位数获取的样本特征同暂未下单的用户行为做**行为相似度**匹配，并计算出暂未下单用户的下单概率。

关键词： 三维可视化 ，随机森林预测模型，行为相似度匹配模型，网格搜索算法，客户消费价值行为

Analysis of the value of user learning product consumption behavior based on random forest and similarity matching algorithm

Abstract: Under the background of the rapid development of Internet, commercial society is closely connected with the Internet, and it has become a common understanding of all industries to combine online channel with offline channel. This requires the use of data mining technology to analyze the user behavior data, build a model to distinguish the value of users, to formulate special marketing strategies for different users, so as to achieve the improvement of user conversion rate with lower marketing cost.The data is preprocessed first to improve the data quality. The data obtained includes four tables, namely, user information table, user login status table, user access statistics table and user order table. Firstly, the data obtained is regulated by Python software. The preprocessing process of data includes processing abnormal samples such as missing value, abnormal index value and index range error to improve the quality of data. Then, the distribution and login of each city of the user are analyzed, and the results are visualized. This paper presents the distribution of all users by using regional map, 3D pentagonal map and three-dimensional thermal map. The conclusion that the single rate of first tier cities is higher is obtained by SPSS software analysis. The user login situation is analyzed by using Excel and SPSS software, and the method of combining the overall analysis with the analysis of the Bureau part is adopted. In the overall analysis, the frequency of user login is analyzed, and the login status of the users who have been placed and not ordered is analyzed in the local analysis. The model is built to determine whether the user will finally order or order the probability of purchase. In this paper, 70% training sample set and 30% test sample set are divided into pre-processing data, and the random forest model is constructed. The importance of each feature is given. The model parameters are optimized by grid search. The model with the highest average accuracy of 98.61% under cross validation is selected as the final model, and the generalization ability reaches 98.64%. The similarity matching model of user behavior is constructed. The sample features obtained by the center number are matched with the behaviors of the users who have not ordered the order, and the probability of the order of the users who have not placed the orders is calculated.

Key words: 3D visualization, Stochastic Forest prediction model, behavior similarity matching model, grid search algorithm, customer consumption value behavior

一、数据预处理

1.1 预处理模型建立

实际上需要分为两步：第一步获取数据，进行数据的规约，即将三张表：用户信息表（user_info.csv），用户登录情况表(login_day.csv)，用户访问统计表(visit_info.csv)进行合并；第二步则是进行数据的预处理，即去除包括存在缺失值、指标数值异常、指标范围错误等异常样本以提升数据的质量。而对于数据预处理的流程图如图 1 所示。

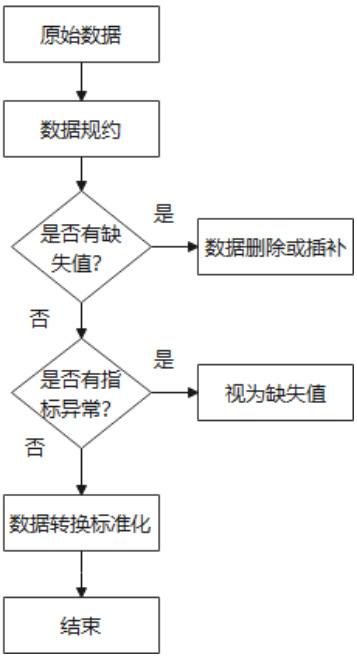

图 1 数据预处理流程图

如上图所示，首先对于原始数据进行数据规约，即完成属性即各项指标的合并。接着进行数值判断，判断样本当中是否存在缺失值，如果存在则进行数据的删除。然后判断是否存在指标异常，如果存在则视为缺失值。最后对数据进行了标准化的处理。

1.2 预处理模型求解

（1）数据归一化处理：原始数据中指标有次数、天数和课程节数的统计，存在不同的量纲单位，因此数据的范围不同，会对预测结果产生影响。归一化的目的不仅是为了展示数据处理后的结果，更是为了归纳统一样本的统计分布性，统一建模和计算之前基本度量单位。

而对于数据的归一化往往有很多计算方法，本文采用的是最小最大法进行数据的归一化的处理，对于所给出的所有指标均采用如下归一化的公式：

$$x_i^{'} = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}} \quad (1)$$

其中 i=[1,47]，$x_i^{'}$ 表示归一化后的第 i 个指标数据，x_i 表示归一化之前的第 i 个数据，$x_{\max}$、$x_{\min}$ 表示第 i 个指标数据下的最大值和最小值。

（2）数值型数据预处理后的可视化展示： 通过程序将不符合范围的、缺失值的样本进行剔除后，并进行归一化之后，再对数据进行可视化展示。由于指标数量众多，这里选择最为典型的预处理后的指标展示，即用户登陆天数指标如图 2 所示。

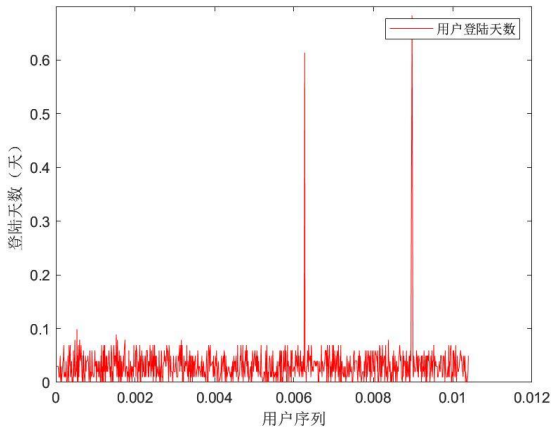

图 2 预处理后的用户登陆天数

如上图所示，通过 python 编程，经过程序处理之后，用户登陆天数、用户登陆间隔、用户登陆距离期末考试天数三个指标的数据均已经符合正常范围（>=0），其他数值型的数据处理方式也是如此。

（3）分类型数据预处理后的可视化展示:分类型数据的处理和展示同数值型数据不同，分类型指标如 chinese_subscribe_num（关注公众号 1）、math_subscibe_num（关注公众号 2）、add_friend(添加销售好友)、add_group(进

群）共同点在于只能取值 0 或 1，因此对于这类指标样本是否存在缺失值以及是否超过定义域的最好检测方法是进行小波分析。利用 matlab 软件对上述四个典型的分类型指标进行小波分析，其他指标处理方法相同，如图 3 所示。

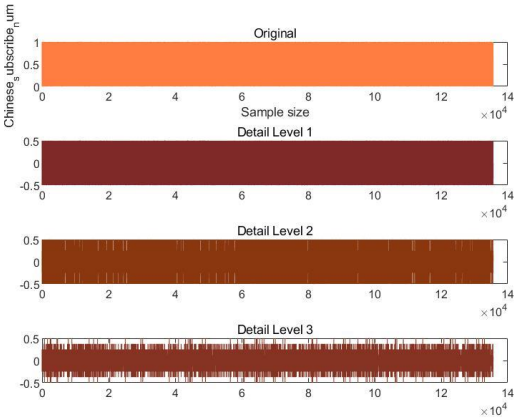

图 3 分类型数据的小波分析

如上图所示，对于一维离散小波进行了加载并重构了 1、2、3 级的细节。从图上就可以看出其处理过的数据加载后，以 chinese_subscribe_num（关注公众号 1）这个指标为例，所有数据均是 0 或 1，并且可以看出这个指标的数据分布非常均匀，经过 1、2、3 级重构后与原数据保持一致，从而说明该指标没有异常点。最后程序给出了异常样本点的位置。

二、用户城市登陆和分布情况分析

2.1 用户城市分布分析

基于上文的数据预处理之后，对用户所在的城市通过 python 软件进行了相关的统计性描述，如表 1 所示。

表 1 用户城市分布统计表

城市	用户数量	城市	用户下单数量
重庆	11880	重庆	326
运城	3459	北京	182
成都	3444	广州	169
广州	2979	深圳	126
北京	2410	上海	112
洛阳	2364	成都	101
保定	1995	东莞	67
泉州	1816	贵阳	65
深圳	1752	佛山	55
邯郸	1559	福州	52
西安	1544	西安	50
郑州	1536	杭州	47
上海	1533	郑州	46
东莞	1417	苏州	43
衡阳	1257	天津	42
福州	1208	衡阳	42
贵阳	1117	南京	36
三门峡	1064	泉州	34
杭州	1041	洛阳	30
佛山	1040	长沙	30

实际上，用户分布的城市多达 361 种，这里仅做部分表格展示。从上表可以看出，**重庆市**用户数目最多，总用户多达 11880 人，其中下单用户为 326 人。除此以外，运城、成都、广州分别包揽了总用户人数榜的前四名，分别为 3459 人、3444 人、2979 人。而下单人数最多的前四名除重庆以外还有北京、广州、深圳，分别为 182、169、126 人。由此可以了解，中国的南方用户数目更多，而一线城市比如北京、广州、深圳、上海等城市下单人数更多，更具有投资推广价值。

接着利用软件对总用户的城市分布数据进行了可视化的展示如图 4 所示。

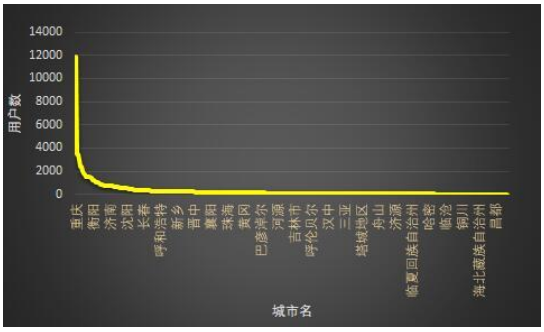

图 4 用户城市分布数据可视化

从上图可以看出，用户城市的分布呈现 L 型分布，对于一些发达一些的一线城市用户不仅数目更多，下单也相对较多；而对于一些稍微落后的地区，用户数量居中，但用户下单较少；最后对于较为落后的地区，用户不仅数量很少，下单的也很少。从树状图也能看出，重庆的用户数目是运城、成都、广州的总和，而一些小城

市的用户数目很少且相差不大。为了更加直观的分析问题，利用 SPSS 软件生成相关图例，用户在各个城市的具体的下单比例如图 5 所示。

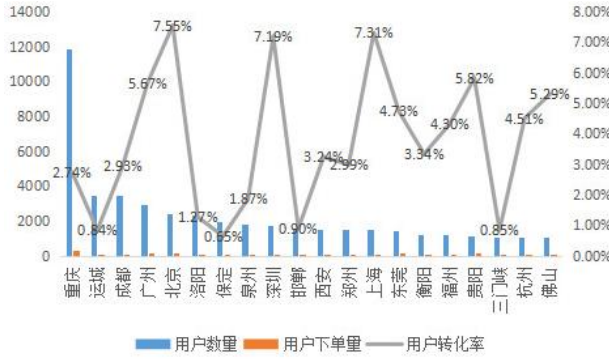

图 5 用户下单率图

从上图中可以直观的看出，在用户分布较多的城市中比如重庆、运城、成都等虽然用户数目很多，但下单比率却相对较小，分别只有 2.74%、0.84%、2.93%；而在北京、上海、广州、深圳这四个一线城市其人数虽然没有前面的城市的多，但其下单率却分别达到了 7.55%、7.31%、5.67%、7.19%。由此可以看出，下单率实际上和城市经济也具有一定的关系，同时该公司的产品质量不错但在一线城市中宣传和吸引程度还是有所不足。

最后，对于用户城市分布情况，利用百度、高德 map 技术进行了二维可视化和三维可视化的展示，用户分布的区域地图如图 6 所示、用户分布三维热力图如图 7 所示。

图 6 用户分布二维区域地图

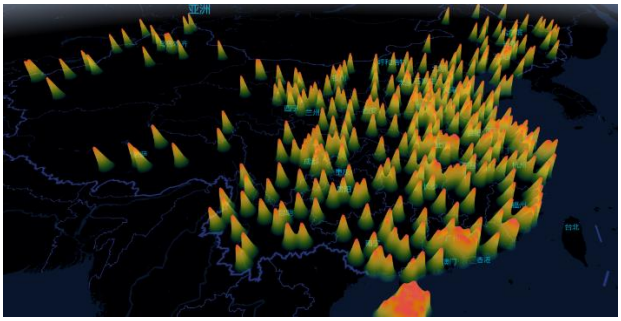

图 7 用户分布三维热力地图

从区域地图可以看出，用户使用已经覆盖全国各地，说明产品的覆盖率已经达到国内要求。但从热力地图上看到，用户大多分布在沿海城市，尤其是南方城市，对于北方城市和西部城市来说，用户数目很少。具体可以通过三维五角地图如图 8 观察。

图 8 用户分布三维五角地图

其中对于三维五角地图，其高度代表了该地的用户数目。从上图可以清晰的看出，各地用户分布不均匀，对于西部地区几乎没有用户。

2.2 用户登陆情况分析

对于用户登陆情况，主要采用整体分析同局部分析相结合的方式。

首先是整体分析，对于所有用户来说，用户登陆情况主要和 login_day（登陆天数）等指标相关。因此先用 python 软件对四个指标进行描述性分析，其频数分析图如图 9 所示。

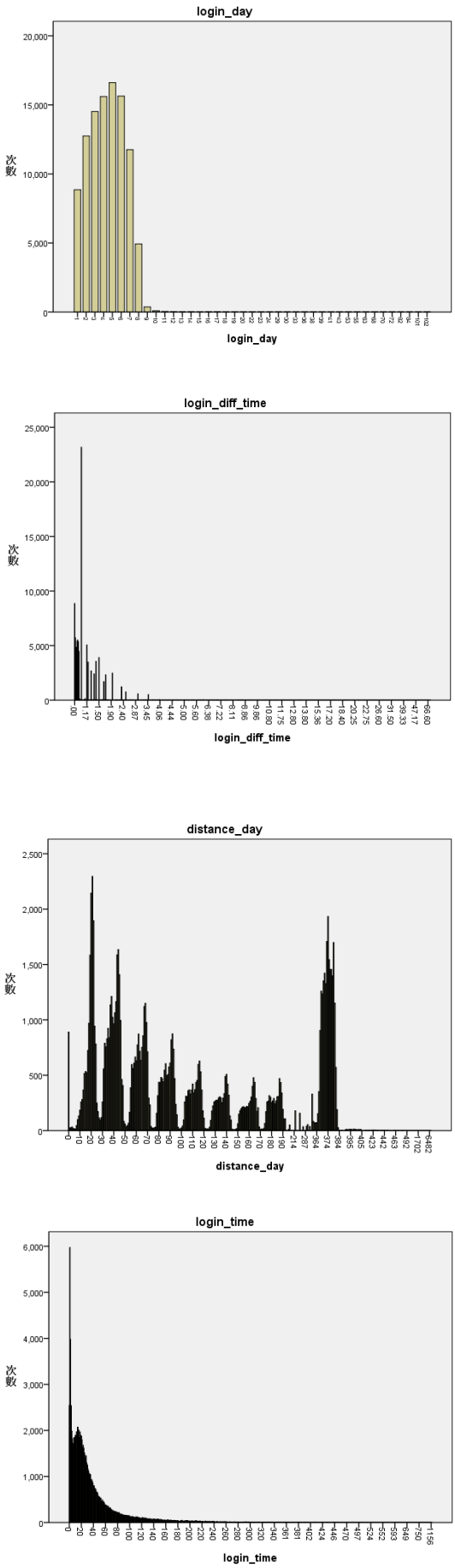

图 9 用户登陆情况频数分析图

上表和上图反映了用户登录天数、登录间隔、最后登录距期末天数和登录时长的平均水平、差异程度和分布形态。从登录情况的描述性统计来看，登录天数的均值在四天左右，最少登录天数为 1 天，最多登录天数为 102 天。用户平均每间隔一天登录一次，登录总时长的均值为 39 小时。离散程度最大的是最后登录距期末天数，说明用户登录在距期末天数方面差异性较大，不具有稳定性。

在分布形态方面，作出四个变量的频数分析图。登录天数、登录间隔和登录时长三个变量都呈现右偏、尖峰的状态。最后登录距期末天数的分布形态在某段时间内具有一定的周期性，可能的原因是日常对于学生会有阶段性的考试，学生会在考试前登录学习网站进行复习，而在考试后用户登录量明显回落。

基于上述分析，可以看出大部分学生一般登陆学习产品主要是因为应付考试，因此相关互联网学习产品在宣传时候应该积极的突出产品在学习上的帮助，从而间接说明产品的质量，增强吸引力。

接着进行局部分析，所谓局部分析就是将下单的学生登陆情况同暂未下单的学生进行对比，利用 matlab 软件进行可视化，如图 10、11 所示。

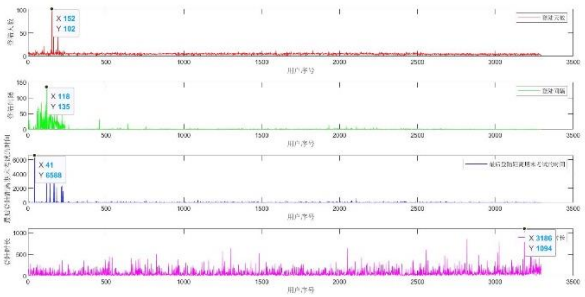

图 10 下单用户登陆情况分析图

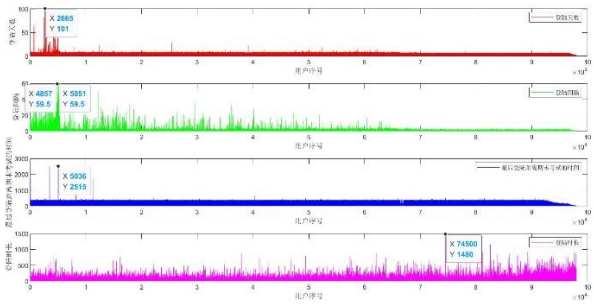

图 11 暂未下单用户登陆情况分析图

从上图中可以看出，在下单用户中，用户 201916804349 登陆天数最多达到 152 天、用户 201721411223 登陆间隔最长达到 118、用户 201595087785 登陆距离期末考试的时间最长达到 6588、用户 202880624847 登陆时长最长达到 1094；而在未下单的用户中，用户 201706256126 登陆天数最多达到 101 天、用户 202298554273 和用户 202319287037 登陆间隔最长达到 59.5、用户 202319223612 登陆距离期末考试最长达到 2515、用户 202723235271 登陆时长最长达到 1480。

从分布上也能看出下单的和暂未下单登陆天数上没有明显差异，均是只有小部分比例能坚持长期登陆学习；在登陆间隔这个指标上，暂未下单的普遍登陆间隔更小，大部分频繁登陆比较活跃，相比之下，下单的反而普遍不活跃；在最后登陆距离期末考试天数这一指标上，暂未下单和下单的保持清一色的一致，即均是考试前突击并“抱团取暖”；而在登陆时长上两者均没有明显差异。

因此基于上述分析，更加启发此类互联网产品的主要价值是在于能够迎合学生突击考试的需求。

三、用户消费行为分析

3.1 模型建立

3.1.1 随机森林预测模型

为判断用户是否下单或者是下单概率，因此考虑将其先转变为二分类问题，并对常用的二分类模型进行选择和改进。基于任务和问题的普适性考虑，首当其冲的选择随机森林法建立二分类模型。而在建立模型之前需要对数据集划分，划分为测试集和数据集。经过异常值检测，处理等预处理后还有 43 个变量，总共 101280 条记录。将前 43 个变量作为输入层，将下单（result）作为输出层，对全部数据按照 7:3 的比例划分训练集和测试集，训练集的样本数为 70896，测试集的样本数为 30384。训练集和测试集的过程中时根据随机参数进行分配，从而避免因人为的主观划分而引起的偏差。随机森林模型的结构流程图如图 12 所示。

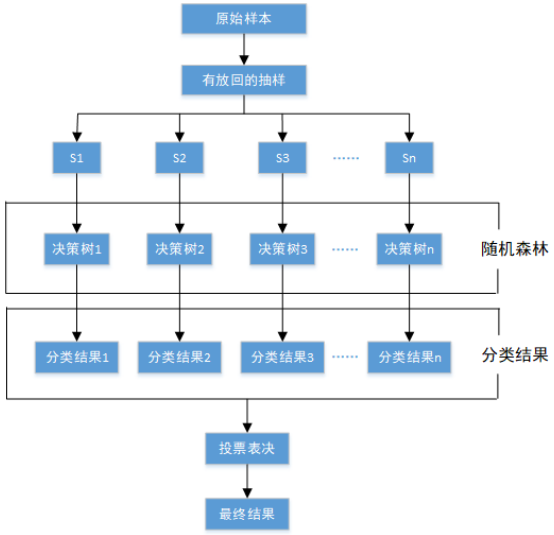

图 12 随机森林算法流程图

随机森林本质上是树型组合分类器$\{h(x,\beta_k),k=1,2,...\}$的集合。通过 CART 算法构建没有剪枝的决策树形成基分类器$h(x,\beta_k)$，其中x是输入向量，β_k为独立同分布的随机向量，共同决定了基分类器的成长过程，最终的输出结果以简单多数投票法则决定[1-6]。

首先对训练集的随机选择。首先通过 bootstrap 抽样算法从原始训练集中抽取 N 个训练子集，每个训练子集的样本数为原始训练集的三分之二，每次的抽样方式都为随机放回抽样。将 N 个训练子集分别成长一棵树从而形成森林，每棵树都不进行任何的剪枝操作，在每棵树成长时通过随机指定 M($M \le N$)个属性参与节点分裂过程，并在这些属性的基础上以最优的分裂方式进行分裂，保证节点分裂的随机性特征。每棵树的分支成长均按照节点不纯度（Gini 系数）最小化的原则（以及其他的评价指标原则），从 M 个属性中选择最优的属性进行分支。不纯度 Gini 系数的公式如下。

$$Gini(S)=1-\sum_{i=1}^{m}p_i^2 \qquad (2)$$

其中C_j为不同类别的分类，S为不同的样本集，P_i为在不同分类出现在样本集中的概率。设$|S|$为样本集S中的样本量，$|S_1|$、$|S_2|$分别为子集S_1和子集S_2中的样本量，则划分的Gini系数的计算公式如下。

$$Gini_{pplit}(S)=\frac{|S_1|}{|S|}Gini(S_1)+\frac{|S_2|}{|S|}Gini(S_2) \quad (3)$$

通过简单投票方式计多为分类结果。最后的输出结果时通过简单的投票方式选择的，通过随机构建多棵决策树进行分类，将每棵树的结果进行简单投票，筛选出票数多的分类结果作为输出值。根据绝对对数投票法的准则得到的公式如下。

$$H(x)=c_{\arg\max_j}\sum_{i=1}^{T}h_i^j(x) \quad (4)$$

其中，h_i是数值型输出值，$H(x)$表示最后投票最多的树的值即最好的树。

3.1.2 行为相似度匹配模型

使用随机森林预测模型解决了用户下单和无法暂定下单的二分类问题，但对于无法暂定下单的用户下单的概率的问题并没有解决。针对无法确定下单的用户，本文创新型的使用用户行为相似度匹配模型，将已下单的用户行为特征与无法确定的用户进行特征相似度匹配，用以代替和预测无法确定用户的下单概率[7-12]。

设i为用户号，j为用户的行为特征，$X_{[i,\ j]}$无法确定用户的行为特征值，$Y_{[i,\ j]}$为已下单用户的行为特征值，θ为相似角度，$M_i=(X_{(i,1)},X_{(i,2)},\dots\dots X_{(i,j)})$为无法暂定下单用户的特征向量，$N_i=(Z_1,Z_2,\dots\dots Z_j)$为已下单用户的特征标准度向量。

对于已下单的用户每个特征，选择每个特征中的点到其他点的距离之和最小的点为下单特征的标准度，即为已下单用户的每个特征的中位数。公式如下：

$$Geometric\ Median=\arg\min\sum_{j=1}^{m}\|Y_j-Y_*\| \quad (5)$$

$$z_j=Y_{[median,\ j]} \quad (6)$$

将无法确定的下单用户的特征组合与已下单用户的标准组合进行遍历匹配，从而计算出相似度如下：

$$cos_i\theta=\frac{M_i.N_i}{|M_i||N_i|}=\frac{X_{(i,1)}Z_1+X_{(i,2)}Z_2+\cdots+X_{(i,j)}Z_j}{\sqrt{\sum_{j=1}^{m}X_{(i,j)}^2}+\sqrt{\sum_{j=1}^{m}Z_{(i,j)}^2}} \quad (7)$$

相似度表示无法确定下单的用户逼近下单的程度，相似度的大小反映了无法确定下单用户的下单概率，利用余弦相似度公式进行计算：

$$P_i=Cos_i\theta \quad (8)$$

3.2 模型求解

3.2.1 随机森林预测模型求解

随机森林预测模型中主要的参数有：最佳的弱学习器迭代次数$n_estimators$，树的最大深度max_depth，内部节点在划分所需最小样本数$min_samples_leaf$，叶子节点最少样本数$min_samples_split$，最大特征数$max_features$。

在实验过程中，本文还将此模型同其他模型算法如逻辑回归模型、支持向量机、决策树模型进行了对比，如图 13 所示。

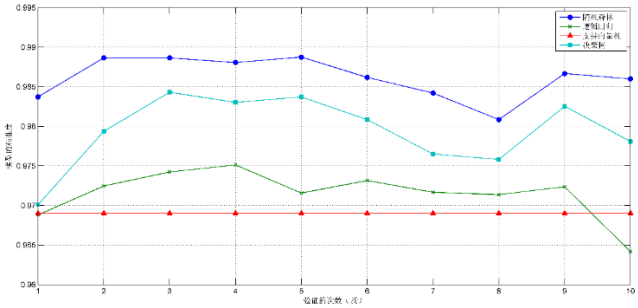

图 13 各个模型的精准度对比

通过各种模型的平均精确度作为所选模型的准则，表 结果显示随机森林预测模型的精准度最高，为98.6167%，其次是决策树，逻辑回归和支持向量机。在模型精确度的稳定性上，随机森林预测模型均优于其他模型，因此在本文的二分类问题中，使用随机森林预测模型是最佳的选择。

3.2.2 行为相似度匹配模型求解

本模型求解核心方法主要采用余弦相似度，通过中心位数获取的样本特征同暂未下单的用户行为做余弦相似度判断，两个用户的行为特征越相似，其下单的概率也就越高，其算法框架如图 14 所示。

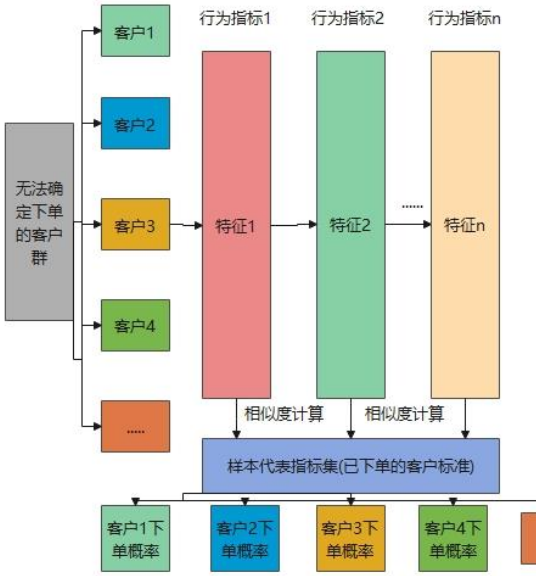

图 14 无法确定下单客户群概率求解框架

由上图可知，通过将用户行为进行特征化后，再选择中心化的样本代表指标集合，再将其与无法确定的下单用户计算相似度用以形成概率。

已下单用户行为特征的标准度，其共同构成的标准向量反映的是已经下单用户的行为标准值，即行为特征里的最优选择。其是通过行为相似匹配算法得出的无法确定下单用户的下单概率，通过与已下单用户各个特征的相似逼近得到其下单的概率估计如表 2 所示。

表 2 无法确定下单用户的下单概率（部分）

用户号	下单概率
用户 1	0.276956
用户 2	0.621672
用户 3	0.737722611
用户 4	0.876129676
用户 5	0.356074881
用户 6	0.443883522
用户 7	0.272854198
用户 8	0.654450121
用户 9	0.454205662
用户 10	0.49807167
用户 11	0.468656866
…	…
用户 49071	0.78968304

3.3 网格搜索算法优化预测模型

3.3.1 网格搜索算法过程

网格搜索通过遍历每个参数组合，通过交叉验证的方式寻找最小 MMSE 来确定最佳参数组合。本文通过 K 折交叉验证对每组的(C,g)的性能指标进行评价。通过 K 折交叉验证的网格搜索流程如下。

（1） 初始化参数的选择范围。令$a=[-a_1,a_2]$、$b=[-b_1,b_2]$,取步长为 1，则网格节点为$C=2^a,g=2^b$。

（2） 样本进行划分。将训练数据进行n等分划分，对网格中的每组(C,g)，任选其中的一个子集作为测试集，其余的$n-1$个子集作为训练集，将所得到的模型对$n-1$等份数据进行预测并且统计均方误差值。

（3） 得到预测值的误差。需要将n个子集都轮流当作测试集，取最后的误差平均值如公式所示：

$$\delta_{MMSE}=\frac{MSE}{n} \tag{9}$$

（4） 寻求最优的参数组合。对参数组(C,g)进行遍历，重复上述的（2）（3）步骤，并且计算在不同参数下的均方误差，对其进行排序，将均方误差最小时对应的的参数组合(C,g)即为最优的参数选择。

结合 K 折交叉验证的方法，设定参数的约束条件，并且以最终的均方误差作为模型搜索参数的选择目标，避免了由于人为主观或者样本随机性所带来的偏差，提高了模型的整体效率与准确率。

3.3.3 优化后的随机森林预测模型

优化后的模型精确度结果如图 15 所示。

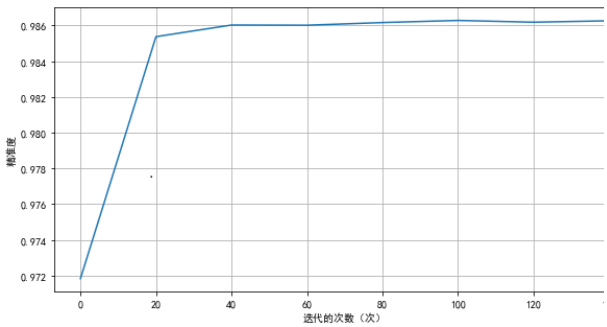

图 15 优化后的预测模型精准度的稳定性

图 16 为RF_1模型 ROC 曲线。ROC 曲线反映的是RF_1模型假正率和真正率之间的关系，X 轴坐标值越接近于 0 代表模型的精准度越高，Y 轴坐标值越大代表模型的精准度越高。AUC 曲线将图形分成两个部分，曲线下面的面积代表 AUC 的值，RF_1模型的 AUC 值为 0.988 非常接近 1，则代表模型的预测准确率极好

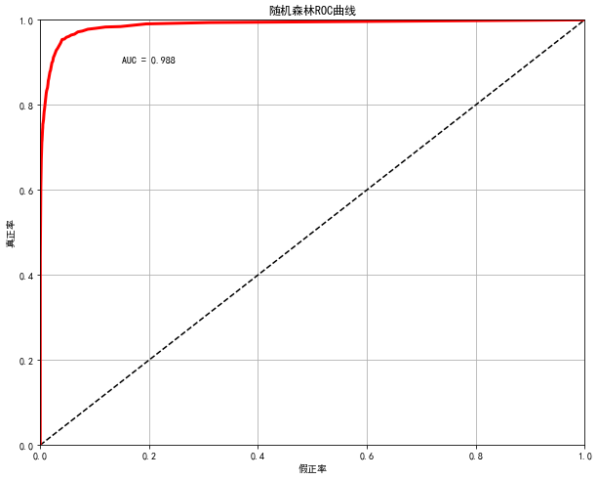

图 16 模型的 ROC 曲线

四、总结

随机森林预测模型的优点：

（1）在处理分类和回归问题时，都可以有较好的表现。由于采用了集成算法，模型在预测准确率的表现上比大多数单个算法要好。

（2）在大样本下，随机森林预测模型在训练速度上具有天然的优势。在子数据集抽样时，采用有放回抽样，有部分数据没有参与到模型的构造中，可以做“袋外估计”，同时减小过拟合风险。在特征维度较高时，随机森林随机选择基学习器节点划分特征，可避免造成“维度灾难”。

（1）随机森林在模型训练的过程中，可以计算特征之间的相互影响，并且可以得出特征的重要性，这对实证结果的解释具有一定的参考意义。

行为相似度匹配模型的优点：

（2）巧妙而又创新性的利用人类行为学原理，把学生之间的行为特征相似度代替学生下单的概率，简化了不确定因素带来的影响。

（3）算法的复杂度较低，适用于处理大批量样本，在大数据行为分析中具有得天独厚的优势。

参考文献

[1] 曹正凤. 随机森林算法优化研究[D]. 首都经济贸易大学, 2014.

[2] 王淑燕, 曹正凤, 陈铭芷. 随机森林在量化选股中的应用研究[J]. 运筹与管理, 25(03), 163-168+177, 2016.

[3] 刘凯. 随机森林自适应特征选择和参数优化算法研究[D]. 长春工业大学, 2018.

[4] 怀听听. 随机森林分类算法的改进及其应用研究[D]. 中国计量大学, 2016.

[5] 曹正凤, 纪宏, 谢邦昌. 使用随机森林算法实现优质股票的选择[J]. 首都经济贸易大学学报, 16(02), 21-27, 2014.

[6] 陈元鹏, 罗明, 彭军还, 王军, 周旭, 李少帅. 基于网格搜索随机森林算法的工矿复垦区土地利用分类[J]. 农业工程学报, 33(14), 250-257+315, 2017.

[7] 王运, 倪静. 基于用户行为序列的概率矩阵分解推荐算法[J]. 小型微型计算机系统, 41(07), 1357-1362, 2020.

[8] 辛磊, 宋玉霞. 社会化商务平台对用户价值共创行为的影响研究[J]. 商业经济研究(24), 59-63, 2019.

[9] 林泽东, 曾庆田, 段华, 鲁法明, 邹杰. 支持活动语义度量的用户行为相似度计算方法[J]. 计算机集成制造系统, 24(07), 1806-1815, 2018.

[10] 赵玮, 李玉萍. 基于消费行为特征的在

线用户价值度量方法实践研究[J]. 商业经济研究(16), 94-96, 2016.
[11]魏逸. 基于概率生成模型的用户行为预测[D]. 上海交通大学, 2015.
[12] T. Chen, J. Chen, k. Zhang, F. Shu and S. Chen, "Research on power consumption behavior analysis based on power big data," 2020 IEEE 9th Joint International Information Technology and Artificial Intelligence Conference (ITAIC), 2020, pp. 591-594, doi: 10.1109/ITAIC49862.2020.9338788.
[13] 罗小燕, 陈慧明, 卢小江, 熊洋. 基于网格搜索与交叉验证的 SVM 磨机负荷预测[J]. 中国测试, 43(01), 132-135+144, 2017.
[14] 纪昌明, 周婷, 向腾飞, 黄海涛. 基于网格搜索和交叉验证的支持向量机在梯级水电系统隐随机调度中的应用[J]. 电力自动化设备, , 34(03), 125-131, 2014.
[15] H. Wang, X. Han, D. Kuang and Z. Hu, "The Influence Factors on Young Consumers' Green Purchase Behavior: Perspective Based on Theory of Consumption Value," 2018 Portland International Conference on Management of Engineering and Technology (PICMET), 2018, pp. 1-5, doi: 10.23919/PICMET.2018.8481949.