Statistics and Application 统计学与应用, 2022, 11(2), 225-233
Published Online April 2022 in Hans. http://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2022.112024

Hans汉斯

基于机器学习对在线教育用户行为的预测

张帅帅

燕山大学，河北 秦皇岛

收稿日期：2022年3月6日；录用日期：2022年3月31日；发布日期：2022年4月7日

摘　要

早在上世纪在线教育就开始在我国崭露头角，发展初期在我国受到各种制约，认可度并不高。然而随着网络不断发展技术不断完善，在线教育发展迅速，目前越来越多人们开始接受在线教育。不只是学生，大学生和工作人群更是在线教育的主要人群。因此网络资源不断增加，各种免费和付费资源层出不穷，很多付费app发现了生财之道，收集有效信息，提高用户对有效知识的接受度。然而，如何找出购买欲望强烈、更有价值的用户，针对性营销，以实现小成本下提升用户转化率是目前互联网普遍面临的问题。本文通过对用户的行为数据进行分析，来挖掘高质量用户所具有的特征，从而帮助企业节省成本，提升利润。针对预处理后的数据集，本文进行了逻辑回归，随机森林预测，XGBoost预测以及LightGBM预测对用户购买行为进行预测，XGBoost以及LightGBM的预测结果相对较好，因此本文是基于XGBoost的预测结果训练和预测的结果对企业提出建议，以提升用户的转化率，增加企业的收入。

关键词

用户价值分析，用户转化率，XGBoost模型，特征重要性，机器学习

Prediction of Online Education User Behavior Based on Machine Learning

Shuaishuai Zhang

Yanshan University, Qinhuangdao Hebei

Received: Mar. 6th, 2022; accepted: Mar. 31st, 2022; published: Apr. 7th, 2022

Abstract

As early as the last century, online education began to emerge in my country. In the early stage of development, it was subject to various constraints and the recognition was not high. However,

文章引用: 张帅帅. 基于机器学习对在线教育用户行为的预测[J]. 统计学与应用, 2022, 11(2): 225-233.
DOI: 10.12677/sa.2022.112024

with the continuous development of the Internet, the continuous improvement of technology and the rapid development of online education, more and more people are beginning to accept online education. Not just students, college students and working people are the main population of online education. Therefore, the network resources continue to increase, and various free and paid resources emerge in an endless stream. Many paid apps have discovered the way to make money, collect effective information, and improve users' acceptance of effective knowledge. However, how to find out more valuable users with strong purchasing desires and target marketing to improve user conversion rate at a low cost is a common problem faced by the Internet at present. This paper analyzes the behavior data of users to mine the characteristics of high-quality users, so as to help enterprises save costs and increase profits. For the preprocessed data set, this paper uses logistic regression, random forest prediction, XGBoost prediction and LightGBM prediction to predict user purchase behavior. The prediction results of XGBoost and LightGBM are relatively good, so this paper is based on the prediction results of XGBoost. The predicted results make recommendations to the enterprise to improve the conversion rate of users and increase the revenue of the enterprise.

Keywords

User Value Analysis, User Conversion Rate, XGBoost Model, Feature Importance, Machine Learning

1. 研究背景

早在上世纪 90 年代，我国互联网刚刚兴起的时候，就已经开始发展早期在线教育了。在模式上，在线教育打破了传统教育的模式，并且在时间和空间上，不再受到约束，学生可以在适合的时间不同的空间一起学习知识。因此，互联网成就了在线教育，在线教育成为了当时互联网创业的热门领域[1]。

发展至今，网络资源数不胜数，我要自学网，哔哩哔哩网站都可以实现网络学习，资源多伴随而来的就是杂质多，选择难，所以付费课程会成为一种高效的选择，付出少量金钱换取时间，更快地学习到有用的知识[2]。本文对使用过某付费课程 app 的用户行为对是否购买课程做用户行为的预测，旨在提高推广效率，减少推广成本，增加收入。

2. 数据说明

从本次课题选择的案例情况来看，在线学习的总样本量为 135,968，购买用户为 4639。比例约为 3%，从这些购买用户当中提取到的总特征数为 40，对这些用户行为序列采取编码的方法开展预测，每条样本由多个用户的行为组成，引入相关的动态属性，对于输入的正向数据需要进行两次编码，选择数值编码器，对输入的信息进行降维取用不同的编码器生成的图像划分为静态和动态两个类型，利用集成学习器开展图像的位置信息，归纳行程表征向量，并进入到下一个机器深入学习翻译的环节当中，输入为句子的顺序单词，提取这些用户行为的序列特征。案例选择所采取的集成学习器建立在 XGBoost 的基础之上，相对于传统的机器学习方法，这种输入输出能够更加独立地描述出用户的行为特征情况，对于购买用户体量较大，达到 4639 人的情况之下，这样的处理是相对合理的。

我们对“login_diff_time”，“login_time”进行了对数变换。原始分布为如图 1 所示：

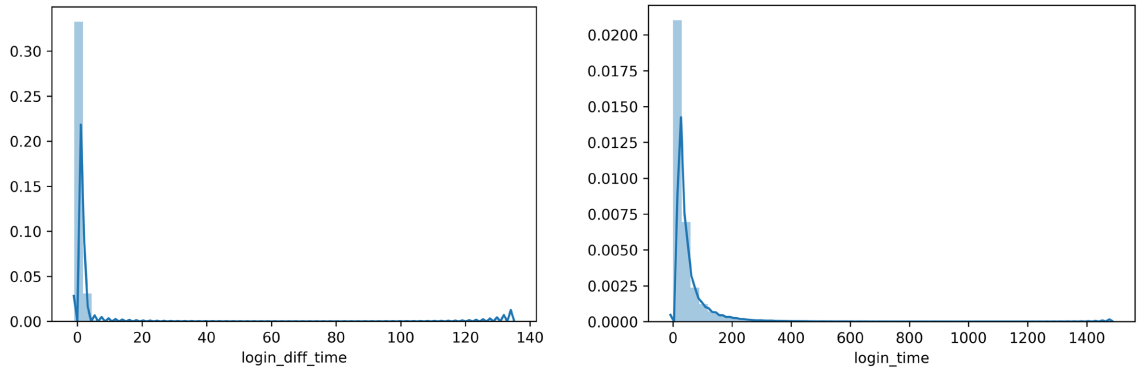

Figure 1. Login_diff_time and login_time distribution
图 1. Login_diff_time 和 login_time 分布

从连续变化向量离散处理结果的形态可以分析得出，0~1 这个范围内，集中了最多的原始向量，对这部分数据开展分析，可以抽象理解为用户的当前兴趣点和短期的行为偏好，让模型既关注当前序列中用户的长期偏好，也关注了短期偏好这两个要提取前后的不同对比，能够较好地展现出模型应用之后在预测效果方面的提升作用。对数变换后分布如图 2 所示：

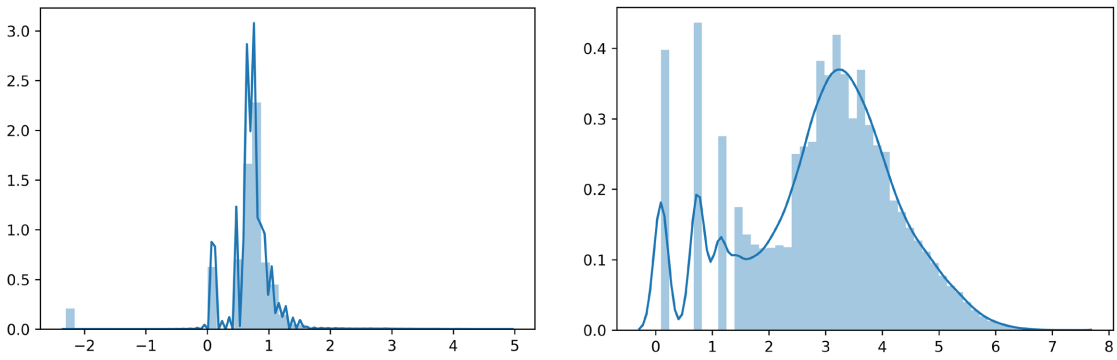

Figure 2. Logarithm distribution of login_diff_time and login_time
图 2. login_diff_time 和 login_time 的对数分布

3. 逻辑回归算法在在线教育用户行为预测中的应用

Logistic 回归属于广义线性模型(generalizedlinear model)。在广义线性模型家族还有多重线性回归。Logistic 回归与多重线性回归大同小异，主要的区别就在于因变量不同，其他方面大抵相同。

广义线性模型家族中模型形式差不多，主要区别在于因变量不同。如果因变量连续，即多重线性回归；如果是二项分布即 Logistic 回归；如果是 Poisson 分布，即 Poisson 回归；如果是负二项分布，即负二项回归。

Logistic 回归的因变量不仅适用于二分类，同时也适用于多分类，其中二分类相对比较常用，也比较容易解释，因此二分类在平时用的比较多，是最常用的 Logistic 回归。

Logistic 回归尽管名字里包括“回归”字样，然而实际上它是一种用于分类的方法，在两分类问题上用途较广(即只有两种输出，分别代表两个类别)，所以利用了 Logistic 函数(或称为 Sigmoid 函数)对于处于线性边界的情况构造预测函数。

逻辑回归是一个非常经典的算法，在机器学习领域有着非常重要的作用，常用于二元分类以及多元分类，在本文中将逻辑回归作为基础学习器的代表，对在线教育的用户行为进行预测，按购买课程人数分层抽样，训练集占比 0.25。得到结果如表 1 所示：

Table 1. Logistic regression prediction results
表 1. 逻辑回归预测结果

	准确率	精确率	召回率	F1	AUC
训练集	0.97	0.45	0.82	0.58	0.96
测试集	0.97	0.41	0.81	0.55	0.95

可以看出，逻辑回归的准确率，召回率和 AUC 指标表现都非常优秀，精确率和 F1 值表现有所欠缺。

4. 随机森林算法在在线教育用户行为预测中的应用

随机森林(Random Forest, RF)是基于许多个树分类器合成的一种分类的算法。由决策树发展的更为高级复杂的随机森林，是多个基分类器集成的 Bagging 算法。RF 的基分类器为决策树。并且值得注意的是，它不单单是使用的决策树算法，还在模型训练时加入了随机的属性选择。详细地说就是在划分属性的选择时，不似以往的决策树算法那样，在运行到当前分支点时，从 c 个属性的集合中自动选择最优的属性。但是就 RF 而言，对以决策树为基分类器的每一个分支点，都会随机地挑出一个拥有 A 个属性的属性集合的子集，再对被挑中的子集选出一个最优的属性来划分下一个分支点。在此，参数 A 代表了这个随机森林模型的随机性程度：当 $A=1$，那么以决策树为基分类器的划分属性标准与普通的决策树是相同的；当 $A!=1$ 那么基分类器属性的划分就充满了随机性。通常来说，推荐 $A=Log_2 c$，此时的效果相对较好。因为是集成 Bagging 算法，以决策树为基分类器的每一个分类器都满足独立的抽样，且森林中的每一个树分类器的随机向量，其值分布相同。在最后的分类时，各个树分类器都会给出一个投票类别，最终以少数服从多数的形式，返回一个结果出现次数最多的类[3]。

随机森林作为集成 Bagging 算法的代表，在用户行为预测中，我们主要从树的最大深度以及树的棵树作为切入点，通过调节这两个参数，使模型达到最好的拟合效果。按购买课程人数分层抽样，训练集占比 0.25。结果如表 2 所示：

Table 2. Random forest prediction results
表 2. 随机森林预测结果

	准确率	精确率	召回率	F1	AUC
训练集	0.98	0.43	0.98	0.6	0.98
测试集	0.97	0.4	0.95	0.57	0.98

当树的深度为 7 且树的棵树达到 300 时拟合效果最好，可以看出与逻辑回归表现相似，在准确率，召回率和 AUC 指标方面表现都非常优秀，精确率和 F1 值表现有所欠缺。

5. XGBoost 算法在在线教育用户行为预测中的应用

XGBoost 是根据 GBDT 基础上改进的算法。GBDT 利用泰勒展开式将目标函数展开到一阶，不同的是，XGBoost 利用泰勒公式将目标函数展开到二阶。在 L_2 正则化项的帮助下，XGBoost 储存了许多目标函数的信息，故而能使模型拥有相对来说更少的损失和更低的方差。XGBoost 和 GBDT 相同，是由 k 个基模型组合而成的一个累加式[4]。

XGBoost 作为集成 Boosting 算法的代表，同随机森林一样，在用户行为预测中，主要从树的最大深度以及树的棵树作为切入点，通过调节这两个参数，使模型达到最好的拟合效果。按购买课程人数分层抽样，训练集占比 0.25。结果如表 3 所示：

Table 3. XGBoost prediction results
表 3. XGBoost 预测结果

	准确率	精确率	召回率	F1	AUC
训练集	0.99	1.0	0.96	0.98	0.99
测试集	0.98	0.84	0.81	0.82	0.99

当树的深度为 5 且树的棵树达到 170 时拟合效果最好，可以看出无论是训练集还是测试集，在各个指标表现都非常优秀，效果远超逻辑回归与随机森林。

6. LightGBM 算法在在线教育用户行为预测中的应用

为了规避 XGBoost 的缺陷，并且能够加快 GBDT 模型的训练速度且不损害准确率，在传统的 GBDT 算法上，LightGBM 做了如下优化：

1) 基于 Histogram (直方图)的决策树算法。

2) 单边梯度采样 Gradient-based One-Side Sampling(GOSS)：使用 GOSS 方法，可以做到只有小梯度且减少大量的数据实例，剩下的具有高梯度的数据用来在计算信息增益的时候使用，在空间和时间上，相比预 XGBoost 遍历所有特征值节省了不少开销。

3) 互斥特征捆绑 Exclusive Fature Bundling (EFB)：为了达到将为的目的，使用 EFB 方法，将许多互斥的特征绑定为一个特征。

4) 带深度限制的 Leaf-wise 的叶子生长策略：因为低效的按层生长(level-wise)不加区分的对待同一层的叶子，所以大多数 GBDT 工具的决策树生长策略使用低效的按层生长，产生了很多没必要的开销。实际上很多叶子的分裂增益较低，没必要进行搜索和分裂。LightGBM 使用了带有深度限制的按叶子生长(leaf-wise)算法。

5) 直接支持类别特征(Categorical Fature)。

6) 支持高效并行。

7) Cache 命中率优化。

LightGBM 针对 XGBoost 的缺陷进行改进，然而实际应用效果仍未可知，本文使用此算法用于在线教育用户行为的预测，通过预测结果可以与 XGBoost 算法作对比，观察其实际应用中的效果[5]。按购买课程人数分层抽样，训练集占比 0.25。LightGBM 预测结果如表 4 所示：

Table 4. LightGBM prediction results
表 4. LightGBM 预测结果

	准确率	精确率	召回率	F1	AUC
训练集	0.99	0.96	0.98	0.97	0.99
测试集	0.98	0.8	0.88	0.84	0.99

当树的深度为 7 且树的棵树达到 500 时拟合效果最好，可以看出无论是训练集还是测试集，在各个指标表现都非常优秀，效果整体与 XGBoost 不相上下且远超逻辑回归与随机森林。

7. 基于评价指标体系四种算法在在线教育用户行为预测中的应用效果对比与分析

从数据的 XGBoost 预测结果分析来看，训练集的准确率达到 0.99，精确率为 1.0，召回率 0.96，F1 值为 0.98，AUC 值为 0.99。从随机森林预测结果来看，准确率 0.98，精确率 0.43，召回率 0.98，F1 值为

0.6，AUC 值为 0.982 者之间在精确率方面存在较大的差异，XGBoost 测试精确率结果要明显高于随机森林预测结果，同时二者在 F1 数值方面也有较大的差异，综合来看，XGBoost 结合随机森林预测，能够弥补随机森林在召回率，精确率，F1 数值方面的不足。Lightgbm 的预测结果与 XGBoost 预测结果之间有较高的统一性，准确率 0.99，精确率 0.96，召回率 0.98，F1 值为 0.97，AUC 值为 0.99。逻辑回归的预测结果与随机森林预测结果有较高的相似性，准确率 AUC，与其他几种方法相似，但精确率值为 0.45。F1 数值为 0.58。从这一数据分析结果可以看出，集成学习器的应用能够对用户的长期行为偏好和短期行为偏好进行一个较为全面的捕捉，这是一种较为有效的基于机器学习的用户集成模型，采取嵌入向量匹配的方法和原理，对四种算法进行基于机器学习的用户偏好行为模型分析，结合迁移的应用，能够减少单一学习器造成的冗余信息过多和预测结果混乱的情况，提高数据的实时性和有效性。

综合来看，优惠券数量是影响用户行为的主要特征之一，一个具体的例子是在一些特殊的节日发放优惠券，更受用户的关注。同时最后登陆距离期末的天数重复课程，中文公众号等相关的因素也会明显揭示用户的行为决策特征，因此需要充分利用这些特征来预测用户的下一个决策。不同的数据信息之间需要考虑到相互补充的功能，利用更丰富的行为数据，解决数据的时效性问题，在行为预测的未来发展情况之下，充分提出解决方案。

特征重要性如图 3 所示：

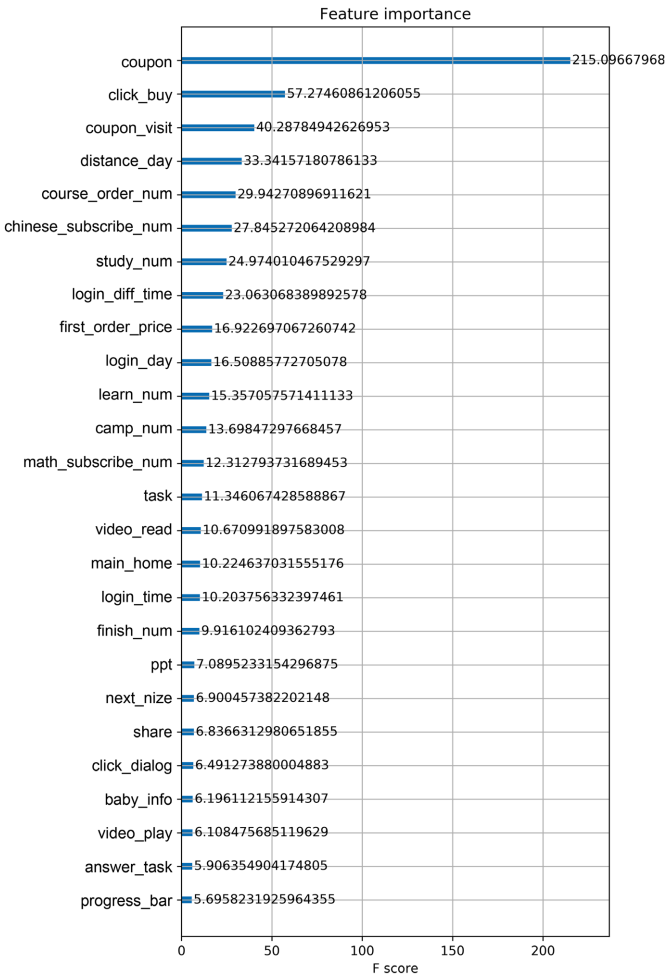

Figure 3. Feature importance
图 3. 特征重要性

可以看出优惠券数量、最后登录距离期末天数、重复课程数、中文公众号等特征对预测影响很大。

8. 基于 XGBoost 算法做模型融合

首先采取的第一个模型融合方式是 XGBOOST + 逻辑回归，简单来讲就是通过 XGBOOST 模型训练得到的树构造新的特征，通过逻辑回归模型对这些新的特征进行重新训练，得到新的训练结果。第二个模型融合是 XGBOOST + 逻辑回归 + 原始特征，很明显就是构造新的特征之后再加入原始特征进行训练。第三个模型是 XGBOOST + bp 神经网络，最后一个是 XGBOOST + bp 神经网络 + 原始特征。得到结果如表 5 所示：

Table 5. Model fusion effect
表 5. 模型融合效果

模型	F1 值	精确率	准确率
Xgb	0.8418	0.7982	0.9897
Xgb + lr	0.8456	0.8051	0.9899
Xgb + lr + 原始	0.8477	0.8086	0.9900
Xgb + bp	0.8297	0.7982	0.9888
Xgb + bp + 原始	0.8371	0.8284	0.9889

可以看出，XGBOOST 融合逻辑回归加原始特征效果更好，在此数据集基础上，融合深度学习效果并不理想，因此模型融合可以达到更好的效果。从另一方面来看，加入原始特征比不加原始特征训练效果更好。因此，在用户行为预测方面，可以采取模型融合且增加原始特征的方式进行预测，提高预测精确率，使广告投放更精准，更少的投资产生更大的收益。

9. 基于结果分析对比对在线教育企业提出建议

用户价值预测模型搭建了用户行为和用户购买预测体系，通过用户行为精准预测了用户是否有购买行为。这样在用户推广方面，可以做到精准定位有价值的用户群体，减少推广成本。为了增加用户购买率以及精准营销，必须首先挖掘用户潜在不购买的原因，从模型的角度输出影响用户价值的重要影响因子，对这些因素做出相应的改善，增加用户购买欲望。

首先从评价体系看模型的优劣性：

从预测准确率来说，XGBoost 和 LightGBM 的训练集和测试集分别为 0.99 和 0.98，其次是随机森林分别为 0.98 和 0.97，逻辑回归为 0.97 和 0.97。可以看出在准确率方面，XGBoost 和 LightGBM 表现最好。

从预测精准率来说，XGBoost 的训练集和测试集分别为 1.0 和 0.84，其次是 LightGBM 分别为 0.96 和 0.8，逻辑回归为 0.45 和 0.41，随机森林分别为 0.43 和 0.4。可以看出在精准率方面，XGBoost 表现最好，LightGBM 表现稍差，随机森林和逻辑回归表现较弱。

从预测召回率来说，随机森林的训练集和测试集分别为 0.98 和 0.95，其次是 LightGBM 分别为 0.98 和 0.88，XGBoost 分别为 0.96 和 0.81，逻辑回归为 0.82 和 0.81。可以看出在召回率方面，随机森林表现最好，LightGBM 和 XGBoost 表现稍差，逻辑回归表现较弱。

从预测 F1 值来说，XGBoost 的训练集和测试集分别为 0.98 和 0.82，其次是 LightGBM 分别为 0.97 和 0.84，随机森林分别为 0.6 和 0.57，逻辑回归为 0.58 和 0.55。可以看出在召回率方面，LightGBM 和 XGBoost 表现最好，随机森林和逻辑回归表现较弱。

对于 AUC，整体差距不大，LightGBM 和 XGBoost 仍占优势。

整体来看 LightGBM 和 XGBoost 预测准确率占相对优势。根据 XGBoost 训练和预测的结果对特征进行分析。

首先，由于互联网社会对于每个人来说获取信息的途径大体一致，然而参与体验课程的用户，大部分用户主要集中在少部分城市中，说明这部分城市可以作为主要发展对象；其次，用户转化率最高的城市分别为北京，深圳，上海，贵阳，广州，佛山、杭州、东莞、衡阳、福州等，主要为一线和新一线城市。通过城市等级划分之后，可以看出：用户主要分布在新一线城市和三、四线城市，但转化率不高；一线城市用户数最低，购买用户数第三，但转化率最高；二线城市用户数第三，转化率第二；一、二线城市还是有很大发展空间。

分析结论如下：

1) 用户转化率最高的三个城市为：北京(7.84%)，深圳(7.79%)，上海(7.28%)。可继续重点在该三个城市深入推广。

2) 新一线城市购买课程人数最多，说明他们有更强烈的需求，可重点关注新一线城市。

3) 购买人群登录天数 90%以上在 8 天以内；登录间隔 90%以上在 2 天以内。最后登录距期末天数 90%以上在 60 天以内。60%的关注了中文公众号；

4) LightGBM，XGBoost，随机森林和逻辑回归对模型预测，LightGBM 和 XGBoost 预测准确率占优势。

5) 使用 XGBoost 模型，测试集(占比 0.25)上拟合准确率为 0.98，AUC 为 0.99；特征重要性最高的 top10 为：coupon (优惠券)；distance_day (最后登录距期末天数)；study_num (课程重复数)；chinese_subscribe_num (中文公众号订阅数)；course_order_num (有年课未完成订单数)；camp_num (开课数)；first_order_price (体验课下单价格)；coupon_visit (优惠券浏览次数)；math_subscribe_num (数学公众号订阅数)；learn_num (课程学习数)。

展望未来，基于机器学习的在线教育用户行为预测，需要利用已有的早期数据训练，让网络拟合大量数据，从而进行权重分析，尽可能提升泛化能力，同时在数据部分的远近和拟合结果的收敛上进行更加有效的训练，现有的分析预测结果验证。LightGBM 和 XGBoost 两种机器学习集成器对于数据的分析更加精准，尤其是在 F1 数值和精准度方面具备显著的优势。利用这两项模型对更大体量的数据开展分析，并结合其他集成学习器开展验证，能够应对未来用户预测数据的多元化指标建设。

考虑到用户的特征，建设差异化的推荐路线是较为有效的方法，针对频繁登陆人群，短时登陆人群和间歇登陆人群采取差异化的策略。注重重点用户的重点转化，尤其是在发放优惠券方面，针对不同的人群设置不同的优惠券领取方法以及优惠券面额，在最后登陆距离距期末天数 0~30 天时，增加用户优惠券的领取，点击入口和推荐频繁程度，通过不同的特征组合应用，达到良好的推荐效果，基于特征预测的影响数值，关注用户的基础行为模型，建立更加有效的迁移数据分析方法，需要对用户的数据开展更加细致的处理，在算法的建设上，未来还有较深层次的内容需要不断地探索，建立模型需要考虑到参数的设置，参数的提取以及参数的转化等问题，未来还需要在功能指标的对比层面，注意力机制的效果分析层面，不同解码方法的效果对比层面，以及引入动态属性的效果分析上进行更有力的探讨。

根据人群登录情况以及模型训练情况我们的建议如下：

1) 大部分购买人群登录天数在 8 天以内；应该抓住用户刚注册时的新鲜感，着重推荐。

2) 购买人群集中在登录间隔小于 2 天，对于频繁登录的人群应该着重推荐。

3) 购买人群集中在最后登录距期末天数的 0 至 30 天，在临近期末时频繁登录的用户应该重点推荐。

4) 领券并且最后登录日期距离期末的天数小于等于 30 的转化率为 0.61，可以对这部分用户做重点推荐，超出 30 天的不做重点推荐。

5) 根据模型训练与预测的结果可以看出优惠券数量、最后登录距离期末天数、重复课程数、中文公众号等特征对预测影响很大，可重点关注用户的这些特征。

本文研究的最后对模型进行了进一步优化，通过优化得到了比基础机器学习模型更好的效果，然而此模型对未来数据进行的预测仍需要未来数据进行校验，预测分类的效果还未可知，因此希望可以等到预测结果得以验证的时候，根据实际数据对模型进行进一步优化。

参考文献

[1] 任传雯. 提高线上教学效果的探究[J]. 现代职业教育, 2021(33): 127-129.

[2] 龙女. 基于学习通平台教育心理学线上教学的实践与思考[J]. 现代职业教育, 2021(36): 128-129.

[3] 葛绍林, 叶剑, 何明祥. 基于深度森林的用户购买行为预测模型[J]. 计算机科学, 2019, 46(9): 190-194.

[4] Koehn, D., Lessmann, S. and Schaal, M. (2020) Predicting Online Shopping Behaviour from Clickstream Data using Deep Learning. *Expert Systems with Applications*, **150**, 103-121. https://doi.org/10.1016/j.eswa.2020.113342

[5] 杨建昆, 夏文财. 基于用户行为分析的诈骗电话识别[J]. 计算机系统应用, 2021, 30(8): 311-316.