文章编号：2095-2163(2023)01-0182-04 中图分类号：TP181；G623.8 文献标志码：A

基于 GBDT 算法的游戏销量预测模型研究

徐英卓[1]，郭 博[1]，王六鹏[2]
（1 西安石油大学 计算机学院，西安 710000；2 西安石油大学 石油工程学院，西安 710000）

摘 要：随着网络游戏的快速兴起，精确的游戏销量预测具有较高的商业价值，能够明确各方投资方向，提高收益，形成合作共赢。本文以影响游戏销量的特征数据为样本，建立基于梯度提升决策树（Gradient Boosting Decision Tree，GBDT）算法的游戏销量预测模型；并将 GBDT 模型预测结果与决策树、线性回归、极端随机树进行对比分析。分析表明，本文所建立的游戏销量预测模型较其它预测模型具有较高的拟合优度，预测效果更好，且在预测阶段的计算速度快，在分布稠密的数据集上，泛化能力和表达能力较好。

关键词：游戏销量；预测；梯度提升决策树

Research on game sales forecasting model based on GBDT algorithm

XU Yingzhuo[1], GUO Bo[1], WANG Liupeng[2]
(1 School of Computing, Xi'an Shiyou University, Xi'an 710000, China;
2 School of Petroleum Engineering, Xi'an Shiyou University, Xi'an 710000, China)

【Abstract】 With the rapid rise of online games, accurate prediction of game sales has high commercial value, which can clarify the investment direction of all parties, improve revenue and form win-win cooperation. Therefore, this paper takes the feature data that affect game sales as the sample and establishes a game sales prediction model based on the Gradient Boosting Decision Tree (GBDT) algorithm. The prediction results of GBDT model are compared with Decision Tree, Linear Regression and ExtraTree. The final results show that compared with other prediction models, the established game sales prediction model has higher goodness of fit, better prediction effect, faster calculation speed in the prediction stage, and better generalization and expression ability in dense distribution data sets.

【Key words】 game sales; forecast; Gradient lifting decision tree

0 引 言

游戏的销量是衡量游戏成功与否的重要指标，对游戏的销量做出一个合理、准确的预测，很大程度上能减少投资风险，使投资收益最大化。当前的销量研究中，采用机器学习对销量预测的研究方法有很多，但是在游戏市场还未对游戏的销量进行预测[1]。机器学习中的非线性模型，如随机森林（Random Forest，RF）、极端梯度提升方法（Extreme Gradient Boosting，XGB）和梯度提升决策树（Gradient Boosting Decision Tree，GBDT）等是以决策树为基本模型的集成学习方法，可把单一学习模型有机结合，形成一个统一的模型，从而获得更准确、稳定的预测学习结果。GBDT 作为较为成熟的集成学习算法，能有效降低预测值和真实值的偏差。通过不断拟合上一颗树的残差来提升性能，更注重学习模型的精度，具有高效、预测准确、对原始数据不敏感、模型的可解释性强等优点[2]。

本文采用 GBDT 算法对游戏销量进行建模预测研究，并综合对比决策树、线性回归、极端随机树 3 种经典回归模型的预测性能和结果。

1 数据集描述及处理

1.1 数据集描述

本文研究的目的是对游戏销量做出预测，采用近十年各个游戏平台主流游戏的特征数据，其中包括训练集 18 000 条数据，测试集 7 000 条数据。游

作者简介：徐英卓（1964-），女，硕士，教授，主要研究方向：石油勘探开发领域应用；郭 博（1997-），男，硕士研究生，主要研究方向：智能计算与可视化；王六鹏（1980-），男，博士，副教授，主要研究方向：钻井信息化应用技术。
通讯作者：徐英卓 Email：707416631@qq.com

收稿日期：2022-04-12

戏特征数据主要包括游戏的名称、发行日期、语言、发行商、支持平台、价格、积极评价数量、消极评价数量等 12 种特征数据。其中，特征数据中包含字符型特征和数值型特征，为保持输入模型参数的格式一致，故使用留一法对字符型特征数据进行变量编码，将其转换为数值型数据。

1.2　数据预处理

数据预处理是提高预测结果准确性的先决条件。数据预处理决定了机器学习训练的上限，而算法和模型的预测结果则更大程度的提高机器学习训练的上限[3]。

本次研究收集的数据较为驳杂，存在跨度较大的数据，并且还存在“0”值以及缺失值。所以在使用数据之前，需要对数据中的缺失值和异常值进行处理。此外，由于特征数据具有不同的测量单位，数据之间的数值差距可能会影响模型，因此需要重新进行数据处理，以避免更重要的特征会影响其他特性，同时提高模型的收敛速度[4]。本文采用 min-max 标准化，使得结果映射到[0,1]之间，如式(1)：

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \tag{1}$$

其中 $x_{\max}$ 为样本数据最大值，$x_{\min}$ 为样本数据最小值。

经过上述归一化处理后，原始数据全部转换为无量纲指标的评估值 x^*，即当评估值处于相同的定量水平，可以进行表征输入[5]。

2　基于 GBDT 的游戏销量预测模型的建立

2.1　GBDT 算法描述

梯度提升决策树是一种迭代的决策树算法，由多棵决策树构成的，每个决策树的结果都是通过加法来确定的。GBDT 算法通过每次迭代在降低残差的方向新建一颗决策树，并在此基础上进一步迭代提高预测结果的准确性。GBDT 通过向前分布算法和加法模型来完成学习的优化过程。该算法的主要流程：首先要初始化第一个基学习器，基学习器是一个只有根节点的树；在此基础上，建立 M 个基学习器，并对其求解损耗函数，将其作为残差的估算值；创建一颗回归树 CART 以拟合该残差；通过拟合后的树叶子节点寻找尽量减少损耗的数值；最后，对学习器进行更新[6]。

GBDT 算法步骤：

初始化基学习器 $f_0(x)$，为式(2)

$$f_0(x) = \arg\min_c \sum_{i=1}^{N} L(y_i, c) \tag{2}$$

其中，$L(y_i, c)$ 为损失函数，用于计算真实值与预测值之间的误差，argmin 为确定损失函数值最小时 c 取值的函数。

(1)建立一系列 CART 回归树，在此基础上利用梯度提升技术拟合残差，GBDT 规定将损失值的负梯度作为残差估计值 r_{mi}，为公式(3)

$$r_{\mathrm{mi}} = -\left[\frac{\partial L(y_i, f(x_i))}{\partial f(x_i)}\right] f(x) = f_{m-1}(x) \tag{3}$$

(2)确定残差估计值后，利用 CART 回归树进行拟合，得到第 m 棵树的叶节点区域 R，其中($j = 1, 2, \cdots, J$)，对于每个叶节点区域，确定使对应损失函数最小化的最佳拟合值 C_{mj}，为公式(4)

$$C_{mj} = \arg\min_c \sum_{x_i \in R_{mj}} L(y_i, f_{m-1}(x_i) + c) \tag{4}$$

(3)更新学习器 $f_m(x)$，为公式(5)

$$f_m(x) = f_{m-1}(x) + \sum_{j=1}^{J} c_{mj} I(x \in R_{mj}) \tag{5}$$

其中，I 为学习率。

本文 GBDT 模型采用分位数损失函数，表达式为式(7)

$$L(y, f(x)) = \sum_{y \geqslant f(x)} \theta \mid y - f(x) \mid + \sum_{y < f(x)} (1 - \theta) \mid y - f(x) \mid \tag{7}$$

其中，θ 为分位数。

本文采用对训练集进行无放回抽样的方法，抽样比例 v 为(0，1]。

2.2　建立游戏销量预测模型

2.2.1　游戏销量预测模型建立流程

以游戏平台实际数据为基础，通过对影响游戏销量的相关因素进行分析，结合数学模型得出合适的模型参数，从而建立游戏销量预测模型。通过这种方式建立的游戏销量预测模型不需经历复杂的分析过程，模型建立难度较小，实用性好。在实际数据中，通过对实际数据处理建立模型，所得模型的准确率较高。基于 GBDT 的游戏销量预测模型建立流程如图 1 所示。

(1)样本数据进行特征工程和数据预处理之后，将全部的游戏销量数据集划分为训练集和测试集；

(2)通过已有的数据模型进行分析，再调整模型参数，并对其进行训练。本文模型所设置的参数包括最大迭代次数、学习率、最大特征数、树的最大深度以及子采样等；

(3)将测试集输入到模型中，得出预测结果；

(4)对模型进行评估、对比和分析。

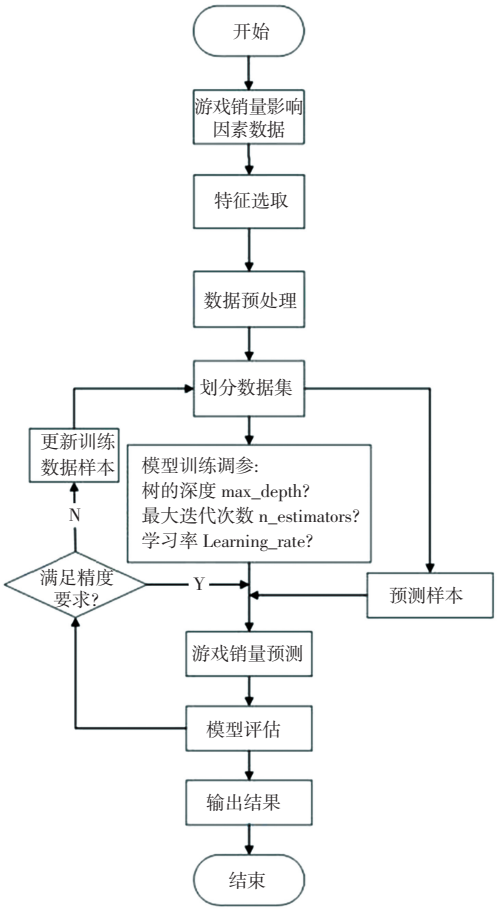

图 1 基于 GBDT 的游戏销量预测模型建立流程

Fig. 1 Process of establishing game sales prediction model based on GBDT

2.2.2 模型参数设置

数据预处理后,对数据进行互信息关联分析,游戏特征参数相关性分析热力图如图 2 所示。

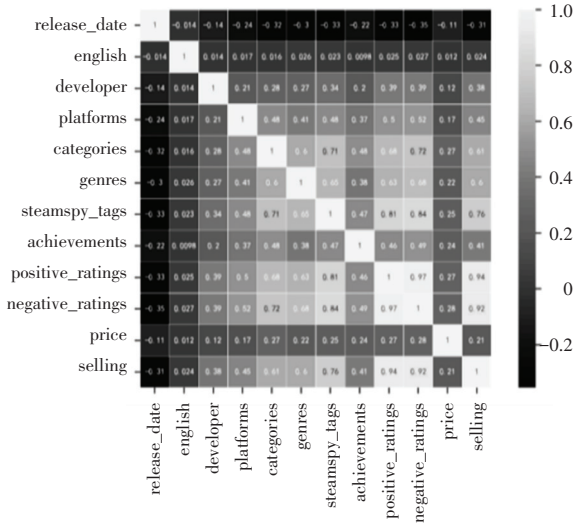

图 2 游戏特征参数相关性分析热力图

Fig. 2 Thermal diagram of correlation analysis of game feature parameters

其中,销量与发行日期,游戏语言互信息值较低,对模型预测无太多参考价值,故舍弃这两个特征。将其他 9 种游戏特征作为游戏销量预测模型的输入变量,建立 GBDT 游戏销量预测模型。同时在实验时使用网络搜索(GridSearchCV)法选择模型的最佳参数,采用五折交叉验证的方法对结果进行验证。游戏销量预测模型的最优参数设置见表 1。

表 1 游戏销量预测模型各参数的含义及取值

Tab. 1 The meanings and values of each parameter of the game sales prediction model

参数含义	值
最大迭代次数	80
学习率	0.1
最大特征数	Sqrt
树的最大深度	8
子采样	0.8

3 实验结果及分析

为了验证 GBDT 算法模型在游戏销量预测的优越性,本文选取决策树、线性回归、极端随机树和 GBDT 优化后模型的拟合优度进行对比分析,按照不同比例划分训练集和测试集,并通过五折交叉验证对结果进行验证。

3.1 实验结果

各个模型的预测结果采用拟合优度(R^2)进行评价,可以直观的观察到各个模型的预测精度,结果见表 2。

表 2 各个模型拟合优度

Tab. 2 Goodness of fit of each model

R-squared	决策树	线性回归	极端树	GBDT
R^2	0.857 1	0.884 3	0.854 6	0.923 2

优化 GBDT 算法后游戏销量预测模型测试集预测结果对比图如图 3 所示,其中因数据量较大,只截取部分数据,便于观察。

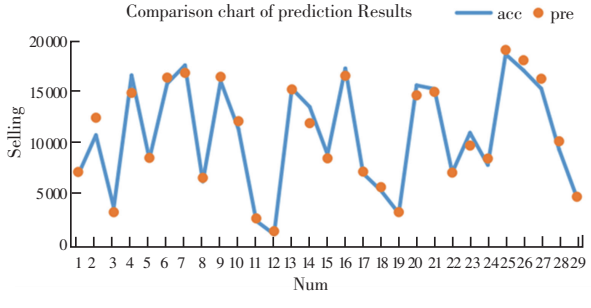

图 3 优化后的游戏销量模型测试集预测结果对比

Fig. 3 Comparison of prediction results of the test set of optimized game sales model

在模型训练中，通过得到各个特征参数的重要性得分，来解释模型的可行性。

计算出每个特征参数的重要性得分，并对其重要程度排序，如图 4 所示。

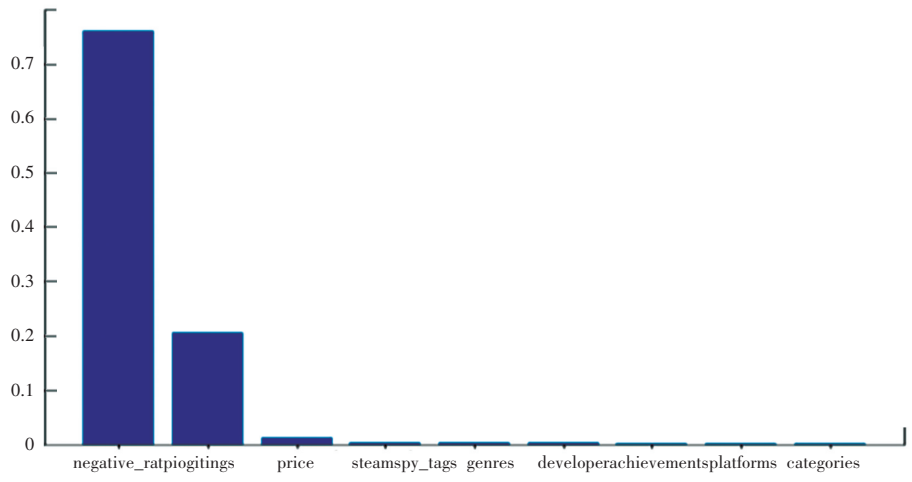

图 4　游戏销量预测模型特征参数重要度排序

Fig. 4　Importance ranking of feature parameters of game sales prediction model

3.2　结果分析

为了分析文中的预测模型的有效性以及预测效果，本文将其与基础预测模型决策树、线性回归和极端随机树进行了对比实验，其中拟合优度 R^2 最大值为 1。R^2 的值越接近于 1，说明当前回归方程对预测值的拟合程度越好。因此本文提出的基于 GBDT 优化的游戏销量预测模型较决策树、线性回归和极端随机树拥有良好的预测精度，可以很好地预测游戏销量，具有较高的可靠性。

通过预测值与真实值对比曲线，可以更为直观的发现分析预测值与真实值的趋势走向以及拟合程度。预测趋势与实际值的趋势比较吻合，但是在拐点处波动较大。

对游戏销量预测模型的特征参数重要度排序，对模型的贡献度最大的特征是消极评价（negative_ratings），其次是积极评价（opstitive_ratings），游戏人数类别（categories）特征的重要性得分最低。消极评价对于销量的影响最为重要，符合结合实际中下载游戏的情况，说明一款游戏的销量，积极和消极的评价起到了至关重要的作用。

4　结束语

（1）应用数据处理对游戏销量预测进行特征工程和信息关联分析，能够有效地去除干扰预测结果的特征，降低噪声干扰和模型冗余，降低其损失值；

（2）基于 GBDT 算法建立的游戏销量预测模型，具有更高的预测精度和准确性，能有效的预测不同特征下的游戏销量、可以为游戏销量提供一定的参考；

（3）本文研究证明了数据驱动模型在游戏销量预测模型应用中的可行性和有效性，为预测游戏销量提供了更为有效的方式和思路。

参考文献

[1] ZHU H, LI H. Predict Prices of Second - hand House Using GBDT Algorithm and PSO Algorithm[J]. Frontiers in Economics and Management, 2021, 2(11): 513-524.
[2] ZHANG W, YU J, ZHAO A, et al. Predictive model of cooling load for ice storage air-conditioning system by using GBDT[J]. Energy Reports, 2021, 7: 1588-1597.
[3] 张鹤. 基于集成学习的蔬菜销售预测[D]. 昆明：云南师范大学，2021.
[4] 吴忆娜，张艺超，袁贞明，等. GRU 和 GBDT 混合模型在早产风险预测中的应用[J]. 计算机系统应用，2022，31(3)：310-317.
[5] 陈陆，吴桦. 基于 GBDT 的船舶油耗预测模型设计[J]. 电子设计工程，2022，30(2)：91-95.
[6] 王小伟. 基于逻辑回归与 GBDT 模型的银行信贷风险评估[D]. 汕头：汕头大学，2021.

（上接第 181 页）

[7] Single Root I/O Virtualization and Sharing 1.1 specification[J/OL]. [2020 - 08 - 03]. https://members. pcisig. com/wg/PCI - SIG/document/download/8238.
[8] 张驰，张傲. SR-IOV 技术在 OpenStack 中的应用[J]. 计算机系统应用，2017，26(9)，246-252.
[9] 汤泳，郭宁. SR-IOV 技术在数据中心网络中的应用[J]. 邮电设计技术，2020，65：65-70.
[10] 冯明振. 基于 macvlan 的 Docker 容器网络系统的设计与实现[D]. 杭州：浙江大学，2016.
[11] 李巍，赵永彬，王鸥，等. 基于 Macvlan 的 Docker 容器网络架构研究[J]. 机械设计与制造，2017，32(5)，270 -272.