上海财经大学

毕 业 论 文

题目

基于机器学习的电子游戏市场的促销策略预测和影响因素研究

姓　　名 李晨茜

学　　号 2018110760

学　　院 统计与管理学院

专　　业 经济统计学

指导教师 周帆

定稿日期 2022 年 4 月 26 日

基于机器学习的电子游戏市场的促销策略预测和影响因素研究

摘 要

随着互联网的发展、电子软硬件技术的进步和消费者对于电子娱乐的需求增大，我国电子游戏市场在过去的十年间飞速扩张。调查表明，营销策略在电子游戏交易环节尤为重要，自 2016 年起，超过 70%的游戏都会在发行后一年内进行有规律的促销活动。而多数消费者更是只会在促销时期购入心仪的产品。这些事实表明，了解电子游戏的促销策略对提升消费者的购买体验尤为重要。

本项目站在电子游戏消费者的角度，以预测电子游戏是否会在特定时间进行促销、探索影响特定游戏促销的特征为目标，建立多元线性回归模型、决策树模型和随机森林模型以达到预测的目的。另外通过分析每个特征所对应模型的 ROC 曲线得到特征的重要性排序，从而达到特征选择的目的。

最终通过模型结果对电子游戏消费者的消费行为提出建议，以期提高消费者的购物体验。

关键字：促销策略、预测、决策树、随机森林、多元线性回归

Research of Forecasting and Influencing Factors on Video Games' Promotion Strategy based on Machine Learning Models

Abstract

With the development of the Internet, the advancement of electronic software and hardware technology, and the increasing consumer demand for electronic entertainment, China's electronic game market has expanded rapidly in the past decade. According to the survey, marketing strategy is particularly important in the video game transaction process. Since 2016, more than 70% of games have had regular sales within one year of release. And most consumers will only buy their favorite products during the sale period. These facts show that understanding the promotional strategies of video games is particularly important to enhance the consumer purchase experience.

From the perspective of video game consumers, this project aims to predict whether video games will be promoted at a specific time, explore the features that affect the sales of specific games, and establish multiple linear regression models, decision tree models, random forest models, GBDT and XGBoost model to get the predicted results. In addition, the importance ranking of features is obtained by analyzing the ROC curve of the model corresponding to each feature, so as to achieve the purpose of feature selection.

Finally, the model results are used to make purchase suggestions to video game consumers, in order to improve the shopping experience of consumers.

Key Words: promotion strategy, prediction, decision tree, random forest, multiple linear regression models

目录

一、引言

1. 项目背景

随着计算机相关技术的融合创新、数字经济的飞速发展和消费者对于精神文化丰富度需求的不断提高，电子游戏产业迎来了自己的发展契机，在不断更新迭代中创造出了属于游戏行业的更大的市场。

在庞大的中国游戏市场中，网页游戏和单机游戏分别占据了游戏市场的半壁江山。相较于网页游戏大多数为免费下载（或不需要下载，仅在网页内就可以运行）、游戏内付费的的营收模式，PC 端游戏则大多以在游玩前进行购买、一次性购买全部内容的模式进行盈利。在中国游戏行业过去的十年间，我国的游戏市场从盗版频生、捆绑下载和游戏质量参差不齐、定价混乱，到了如今不断细分和规范化的市场，有众多游戏公司、平台和发行商在期间起了不可或缺的作用。

2. 文献综述

根据《2021 年中国游戏产业报告》显示，2021 年中国游戏市场销售总额达到 2965.13 亿元，同比增长 6.4%；中国游戏用户规模稳定增长，用户规模超过 6 亿人，同比增长 0.22%。

Steam 游戏平台（中国大陆也称为蒸汽平台）是美国电子游戏商 Valve 于 2003 年 9 月 12 日推出的数字发行平台，提供数字版权管理、多人游戏、流媒体和社交网络服务等功能。Steam 被认为是全世界的电脑游戏界最大的数字发行平台，2021 年，Steam 的每月活跃用户数为 1.32 亿、每日活跃用户数为 6900 万。2016 年，Steam 首次打通国内线上支付渠道，被视为 steam 进军中国市场的起点。2018 年 11 月，美国 Valve 公司在上海浦东与完美世界签订协议，“Steam 中国”正式落户上海浦东，进一步促进了 Steam 在中国市场的发展。

在这些庞大的市场数据背后，我国的游戏消费者的群体在不断壮大，该群体对游戏这一娱乐形式的认知也在不断提高。10 年前的消费者版权意识相对较弱，加之当时的网络上盗版横行，游戏开发商也缺少对已发行游戏的定期更新和维护，用户大多不愿为没有实体的“电子游戏”付款。然而，随着消费者对游戏质量的要求提高和如 Steam 一样的游戏交易、游玩平台的不断发展，如今的电子游戏消费者大多更加愿意为一款优秀的游戏买单。

随着消费者对电子游戏经营促销模式的了解的加深，大多数消费者逐渐在自身的购买经历中学会了等待游戏促销的时机，以期在游戏发售后尽可能短的时间内以尽可能优惠的价格购买到自己喜欢的游戏。然而，个体消费者对游戏促销的判断多来自个人经验，这种经验有时会奏效，有时却会导致用户在新游戏发售后等待过长时间或刚购买完就遇到打折到历史新低价格的情况。

为了最优化用户的购买体验，尽可能在节约消费者的经济资源的同时最大化消费者的消

费体验，使其能预先知晓新发售的游戏的打折促销时间和幅度就变得尤为重要。根据相关研究报告显示，游戏的促销策略通常受发行商营销手段、产品本身特征和产品在市场上的评分等因素影响。这也侧面说明了游戏促销的行为是可以在一定程度上被提前预知，从而优化消费者体验。

二、研究目标与研究意义

1. 研究目标

1） 针对一个新发行的游戏，预测其在发行后一个月内是否会进行促销活动；并探索哪些因素对游戏促销有较为显著的影响

2） 针对一个新发行的游戏，预测其促销幅度超过 50%所需要的时间

3） 针对模型得出的结果对消费者提出购买游戏时的建议

2. 研究意义：

作为 PC 端单机游戏的主要发行及交易平台之一，根据 Steam2021 年度回顾专题报道，Steam 玩家们在 2021 年共游玩了大约 280 亿小时，相较于前年增长了 21%。而 Steam 平台的玩家消费额更是比 2020 年提高了 27%。在 2021 年，Steam 平台每个月都会迎来 260 万首次购买游戏的玩家，其增长率和 2020 年疫情下的新玩家增长率几乎不相上下。

一个老练的 Steam 用户，往往在浏览商店后会将心仪的游戏加入愿望单中，但这并不意味着愿望单中的游戏有朝一日能够被购买、加入玩家的游戏库。折扣太低、错过打折时间等都会导致游戏在愿望单中落灰。而如果是一个新加入 Steam 平台的玩家，他很有可能因为尚不清楚 Steam 打折机制而白白消耗掉大量的自己的钱财。同时，由于很多热门游戏的促销时间并不一定会与 Steam 平台的大促日历完全吻合，用户对心仪游戏打折期的错误预估往往会导致游玩计划落空，带来不佳的娱乐体验。

构建游戏促销预测模型对于提升用户体验，帮助新用户在进入 Steam 平台时或老用户面对一款新发售的游戏时做出购买规划是至关重要的。同时，Steam 作为游戏行业占据极大市场份额的游戏平台，对其营销策略的研究和预测也能帮助我们更好地了解整个游戏行业在决定一款产品是否进行促销和促销力度时最看重的因素是什么。

三、数据说明与变量解释

1. 研究对象与数据说明

- 本文的研究对象为 52,483 个在 Steam 游戏平台发布的电子游戏的相关特征指标

（预测变量）与其发售价格和促销信息（响应变量）

- 游戏的基础信息（如名称、种类、特征、发行日期、发行价格等）来源于 Steam 官方网站（https://store.steampowered.com/）游戏详情页；由于 Steam 官方并不提供游戏价格的历史信息，因此游戏的历史价格信息来源于 IsThereAnyDeal（https://isthereanydeal.com/）网站；由于 Steam 官方并不提供玩家数量和玩家游玩时长的信息，因此关于玩家的信息来源于 SteamSpy（https://steamspy.com/）网站。
- 经整合的初始数据集共含有 20 个变量，52,483 条观测。

2. 变量解释

对预测变量和响应变量进行数据类型展示和变量解读，如表 1 所示：

表格 3-1 变量说明表

变量类别	变量名称	变量解释	附注
游戏基础信息	gamename	游戏名称	
	gameid	游戏 ID	每个游戏的 ID 是独有的
	description	游戏简短介绍	
	taglist	标签	一个游戏可有多个标签，共有 427 种标签
	language	游戏支持的语言	一个游戏可支持多种语言，共有 29 种语言
	system	游戏支持的系统	Win; Mac; Linux
	genre	游戏类别	一个游戏可以有多个类别，共有 29 个种类
	features	游戏特征	一个游戏可以有多个特征，共有 43 种特征
	format_release_date	游戏发行日期	
游戏用户信息	reviewsummary	近 30 天用户评价	共有 6 种取值，None 表示在近 30 天内没有用户评分
	reviewsummary_forever	所有用户评价	共有 6 种取值
	score_forever	所有用户评分	取值为 0-1
	rating_sample_num_forever	所有评分用户数量	
	avg_forever	平均游玩时间	分钟
	avg_2weeks	近两周平均游玩	分钟

		时间	
	median_forever	游玩时间中位数	分钟
	median_2weeks	近两周游玩时间中位数	分钟
	ccu	最大同时游玩人数	取数前一天的数据
游戏价格信息	release_price	发行价格	人民币
	first_discount_date	发行后第一次打折日期	

四、数据预处理

1. 缺失值模式探索

由于 2016 年 steam 首次打通国内线上支付渠道，可以将此视为 Steam 进军中国市场的起点。在此之前 Steam 中国区的销售数据较少且代表性较弱，因此在数据集中只取发行日期晚于 2016-01-01 的游戏相关数据（共删除 8313 个）。在从 IsThereAnyDeal 网站获取游戏历史价格数据时，由于网站中部分游戏历史价格相关数据缺失，导致有 8550 个游戏没有历史价格信息，将它们也从数据集中删去。最终剩余 35620 条数据

下面进行缺失值的检验，结果如下图所示：

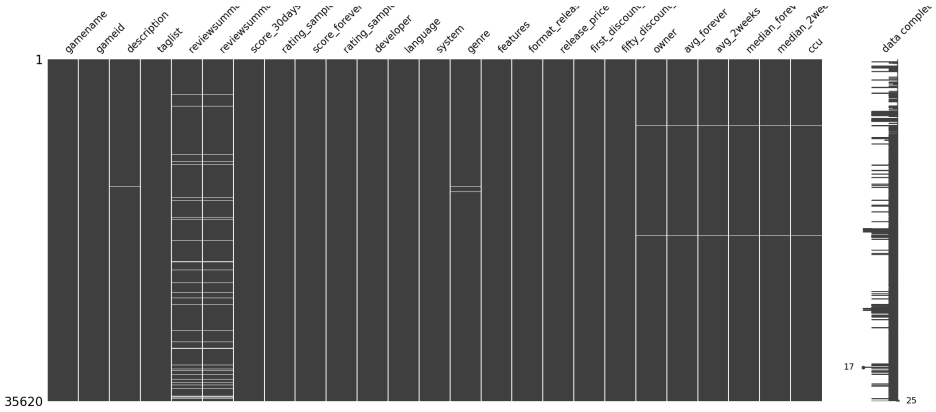

图 4-1 数据集缺失情况 1

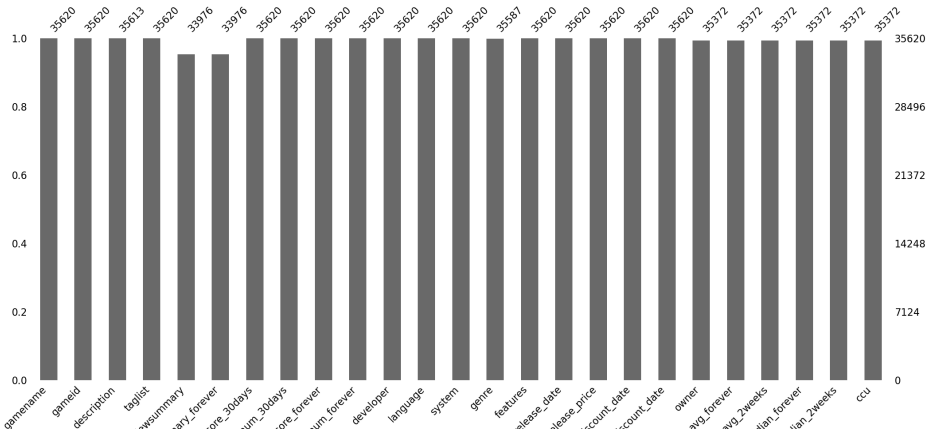

图 4-2 数据集缺失情况 2

数据集共有 35620 条观测，其中，无缺失的完整数据共有 33976 条（如图 2）。可以发现游戏描述（description），近 30 天用户综合评价（reviewsummary），所有时间的用户综合评价（reviewsummary_forever），游戏类别（genre），游戏拥有者数量（owner），平均游玩时间（avg_forever），近两周平均游玩时间（avg_2weeks），游玩时间中位数（median_forever），近两周游玩时间中位数（median_2weeks），同时在线人数峰值（ccu）这 10 个变量均有不同程度的缺失。

从实际意义上来看，每条数据表示的是不同的游戏的客观特征，具有较强的特异性，不太适合进行缺失值插补；缺失的变量大多是游戏的用户数据（如游玩时间、拥有者数量等），且经过查看发现是某几条数据同时出现这样的用户数据缺失。猜测可能是由于游戏过于小众，没有任何购买者，所以没有任何用户使用和评价记录。同时，缺失的数据只占总数据的 5%左右，占比相对较小。结合以上原因，为了尽量保持数据集的完整性并不违背其显示意义，将缺失的变量值用 0 进行替代。

2. 变量合并和选取

在原始的从各网站爬取的变量中有很多变量存在多个取值，如类别（genre）变量可取值范围有 29 种，一个游戏可以同时从属于多个类别，即类别间并非互斥关系。这样的变量没有办法直接输入模型。此外，变量 taglist、features、language 也存在同样的问题。为了解决这个问题，我们根据变量的现实意义和数据占比对变量进行合理的合并和删除整理。

1）**游戏类别（genre）**

数据集中游戏类别（genre）有 29 种可能取到的值，根据每种游戏类别在 35620 条数据的出现频率，可以获取依照频次倒序排列的游戏类别表，如下表所示：

表格 4-1 游戏类别频数频率表

游戏类别	cnt	freq
独立	26780	0.733
动作	15660	0.429
休闲	14870	0.407

冒险	13561	0.371
模拟	6935	0.19
策略	6639	0.182
角色扮演	5239	0.143
抢先体验	4283	0.117
免费开玩	2338	0.064
体育	1833	0.05
竞速	1372	0.038
大型多人在线	877	0.024
实用工具	469	0.013
设计和插画	280	0.008
动画制作和建模	220	0.006
教育	225	0.006
视频制作	167	0.005
音频制作	133	0.004
游戏开发	105	0.003
照片编辑	69	0.002
软件培训	87	0.002
网络出版	53	0.001
财务管理	12	0
电影	1	0
纪录片	1	0
剧集	1	0
短片	1	0
教程	1	0
360 全景视频	1	0

可以发现有一些类别并不属于游戏的范围，如“游戏开发”、“动画制作和建模“、”使用工具“、”财务管理等。并且大部分分类的占比非常低，很多分类只有不足 5%的数据占比。同时，根据游戏分类的现实意义，有一些游戏类别（genre）的取值并不能完全表示游戏种类，例如“抢先体验”和“免费开玩”只是该游戏的一种属性，而并不属于游戏内容的分类。

综合以上原因，最终只保留“独立”、“动作”、“休闲”、“冒险”、“模拟”、“策略”六大游戏分类，并将其分别转化为六个二分类变量。这六大游戏分类可以很好地覆盖绝大多数数据条目，只有 1223 条数据不符合上述的任何一种分类，认为不会对数据整体造成显著的影响。

2）**游戏特征**（features）

数据集中游戏特征可能的取值有 43 种，同样，每个游戏可以拥有多个特征而各个特征之间也不一定存在互斥关系。经过考量现实意义和各种特征的出现频次，删除出现频率低于5%的特征，并将一些表意类似的特征进行合并，最终结果如下：

合并“线上玩家对战“,”在线合作“,”同屏/分屏玩家对战”,“同屏/分屏合作”,“远程同乐”,“跨平台联机游戏”,“大型多人在线”,“局域网玩家对战”为“多人”；合并“完全支持控制器”,“部分支持控制器”,“手柄”为“支持控制器”；合并“在手机上远程畅玩”,“在平板上远程畅玩”,“在电视上远程畅玩”为“远程畅玩。

最终保留“steam 成就”、“steam 创意工坊”、“应用内购买”、“单人”、“steam 排行榜”、“远程畅玩”、“多人”、“支持控制器”总共八个有关游戏特征的变量，并将其分别转化为二分类变量。

3）**用户自定义标签**（taglist）

用户自定义标签是由用户在游玩游戏后、测评游戏时给出的标签，Steam 官方将众多用户个人给出的标签综合选出提出率高的标签作为用户自定义标签。由于用户自定义标签的生成机制，其可能取值有 417 种之多。并且每种标签的占比都相对较小，不存在如游戏类别一样少数几个类别可以覆盖绝大多数游戏的情况。

然而，如果换一个角度考虑，用户自定义标签大约可以类比于用户给游戏的评测。一个游戏的用户自定义标签越多，表明该游戏在玩家中越受欢迎或是游玩该游戏的玩家越多。因此，我们不再关注游戏具体有哪些用户自定义标签，而是关注每个游戏共有多少个用户自定义标签，这个变量不仅能反映游戏的用户广度，更能侧面展示游戏内容的丰富度。

4）**支持语言**（language）

支持语言的取值最多有 29 种，意味着数据集中的一个游戏最多可能支持 29 种语言。由于支持一些相对小众的语言的游戏相对较少，为了简化变量，计数每个游戏支持的语言数量，作为新的变量 language_cnt。同时，选出 5 种在国际上根据使用人数、国家等因素评分最高的语言（英语、中文（简体中文和繁体中文）、俄语、西班牙语、法语）并转化为二分类变量。

5）**其他**

reviewsummary 和 reviewsummary_forever 是根据用户好评度得到的分类变量，根据好评度不同分为如下几档：好评如潮、特别好评、多半好评、好评、褒贬不一、差评、多半差评、特别差评、差评如潮。由于该指标是根据用户的好评率得出的，因此可以将其简化为连续变量，依序为评级赋分，最终这两个变量为 0-9 的评分数。

经过上述变量处理后，数据集中的变量名称及解释更新如下：

表格 4-2 合并/拆分变量后的变量解释表

变量类别	变量名称	变量解释	附注
游戏价格信息	release_price	游戏发行价格	人民币

	first_month_discount	发行后第一个月是否打折	1=打折；0=不打折
	first_discount_percent	发行后第一次打折力度	百分比
	first_discount_period	发行后第一次打折距发售时间的天数	
游戏基本信息	language_cnt	支持语言数量	
	language_en	是否支持英文	1=支持；0=不支持
	language_cn	是否支持中文	1=支持；0=不支持
	language_ru	是否支持俄语	1=支持；0=不支持
	language_ french	是否支持法语	1=支持；0=不支持
	language_spanish	是否支持西班牙语	1=支持；0=不支持
	genre_独立	是否属于独立类别游戏	1=支持；0=不支持
	genre_动作	是否属于动作类别游戏	1=支持；0=不支持
	genre_休闲	是否属于休闲类别游戏	1=支持；0=不支持
	genre_冒险	是否属于冒险类别游戏	1=支持；0=不支持
	genre_模拟	是否属于模拟类别游戏	1=支持；0=不支持
	genre_策略	是否属于策略类别游戏	1=支持；0=不支持
	features_Steam 成就	是否在 Steam 平台上有成就功能	1=支持；0=不支持
	features_Steam 创意工坊	是否在 Steam 平台上有创意工坊	1=支持；0=不支持
	features_应用内购买	是否支持应用内购买	1=支持；0=不支持
	features_单人	是否是单人游戏	1=支持；0=不支持
	features_Steam 排行榜	是否处于 steam 榜单上	1=支持；0=不支持
	features_远程畅玩	是否支持远程畅玩	在电视上；在平板上；在手机上
	features_多人	是否是多人游戏	1=支持；0=不支持
	features_支持控制器	是否支持控制器	1=支持；0=不支持
游戏用户信息	tagsum	用户自定义标签的数量	
	avg_forever	平均游玩时长	分钟
	avg_2weeks	2 周内游玩时长	分钟
	median_forever	游玩时长中位数	分钟

	median_2weeks	2 周内游玩时长中位数	分钟
	ccu	同时在线人数峰值	
	reviewsummary_score	近 30 天用户好评率	
	reviewsummary_forever_score	用户好评率	
	score_forever	用户评分	
	rating_sample_num_forever	评分用户数量	

3. 异常值处理

对于各连续变量画出箱线图并结合现实情况检验异常值。发现异常值出现在游戏的发行价格上（release_price），箱线图如下图所示。根据 Steam 官方历年发布的报告数据和 SteamSpy 可知，Steam 上游戏的平均定价在 0~200 元左右，定价高于 500 元的往往并不是游戏（Steam 上偶尔也有一些动画建模工具等设计软件），而定价高于 1000 元的往往是数据获取的时候出错，因此出现了异常值。因为本文研究的是有关游戏促销的相关问题，并且在数据集中游戏发行价格异常的数据仅有 319 条，占比非常少，因此将该 319 条数据删除。删除异常值之后发行价格的箱线图如下图所示。

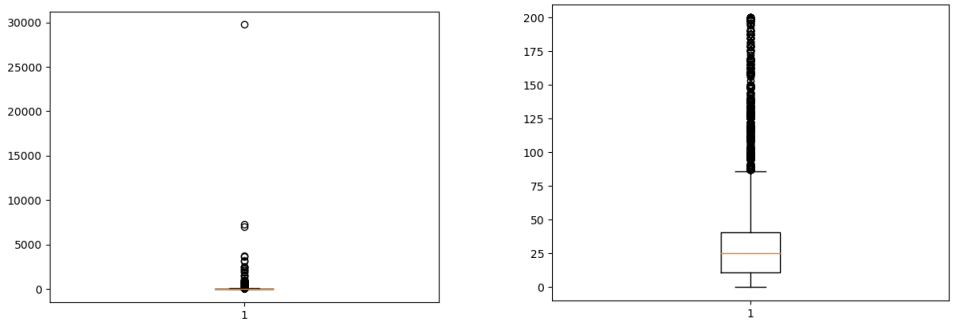

图 4-3 数据集异常值情况

经检验，其他变量中均不存在异常值。

五、描述性统计分析

1. 游戏的基本信息

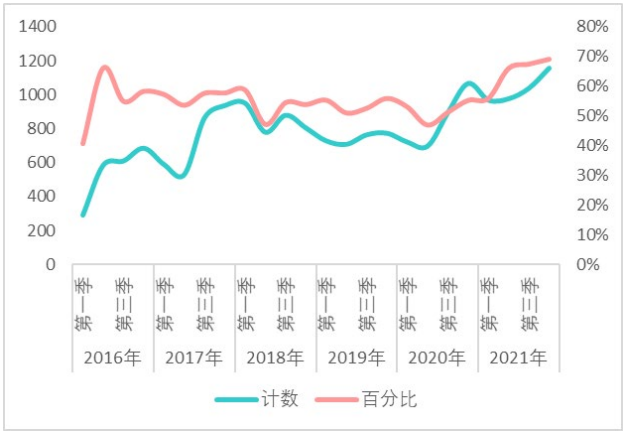

图 5-1 游戏发布数量时间图

首先，按发布时间来看，2016 年之后发布的游戏在第一个月内的打折的数量是在逐年攀升的。然而这也有可能和在 Steam 游戏平台发布的游戏总量越来越多有关系。如果按游戏发行时间分季度计算第一个月打折的游戏数量占比，可以发现该比率在 2016 年之后是基本持平的。

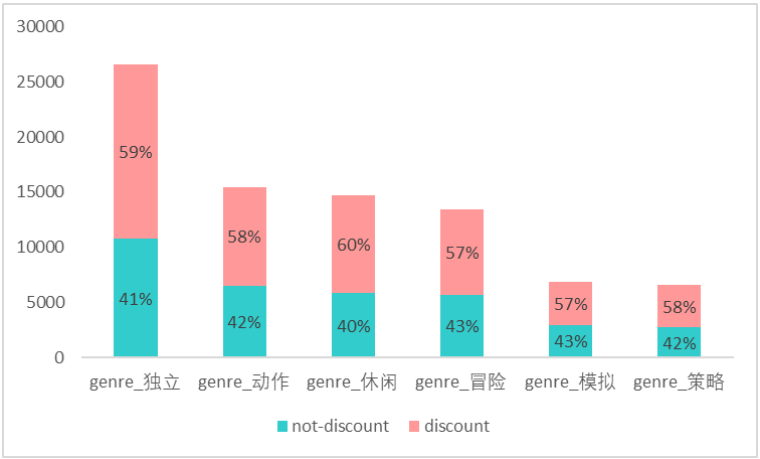

图 5-2 游戏类别与是否促销关系

从游戏类别来看，在该数据集中，独立游戏的数量最多，而策略类游戏的数量则最少。在每一个游戏类别内，发行后第一个月就进行促销活动的游戏占比在 57%~60%不等，休闲类的游戏中在发行后一个月内进行促销的相对略多，而冒险类游戏在发行后一个月内进行促销的相对略少。

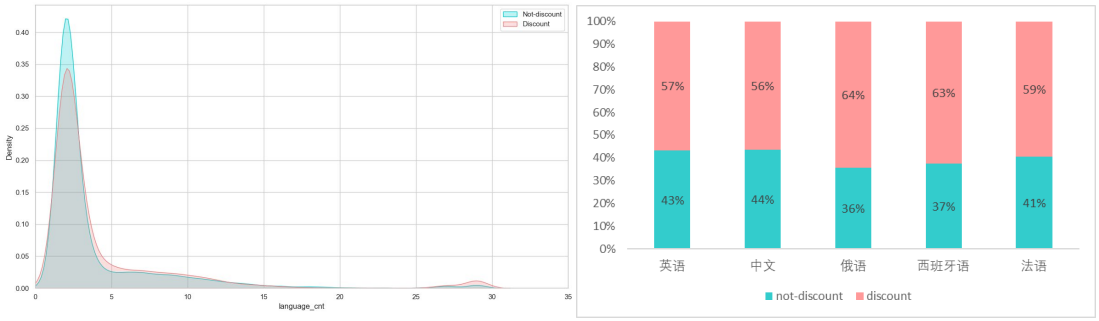

图 5-3 游戏支持语言与促销关系

根据每个游戏支持的语言的总量和对选出的五种主要语言的支持情况不同，可以画出上述密度图和柱状图。从游戏支持的语言的总量来看，绝大多数游戏支持的语言数量在 0~5 种之间，并且随着支持语言数量的增加，游戏数量是有明显减少的趋势。考虑是由于每种游戏的主要受众国家和人群有限，且游戏团队的经济预算往往也不允许每个游戏都支持所有语言。然而，可以发现当支持语言数量提高到 30 种左右时，游戏数量反而有了一定的提升，猜测这部分游戏是因为发布后广受好评，受众面不断增大，因此制作方增加了支持语言的数量。这样优秀的游戏数量理论上应该占比非常小，也和密度图上显示的相符合。

从图中可以看出，在支持 1~2 种语言的游戏中，发行后第一个月进行促销的游戏数量明显小于不进行促销的游戏数量；而随着支持语言类别的增多，发行后第一个月进行促销的游戏数量也在不断提升，直至高于不进行促销的游戏。考虑可能是由于支持语言较多的游戏往往受众群体广，销量高，因此即便促销也可以实行薄利多销的策略来维持发行商的利益。

从主要使用的五种语言来看，游戏发行后第一个月进行促销的比例在 56%~64%之间。其中，支持俄语的游戏在第一个月进行促销的比例最高；而支持中文的游戏促销的比例相对略低。

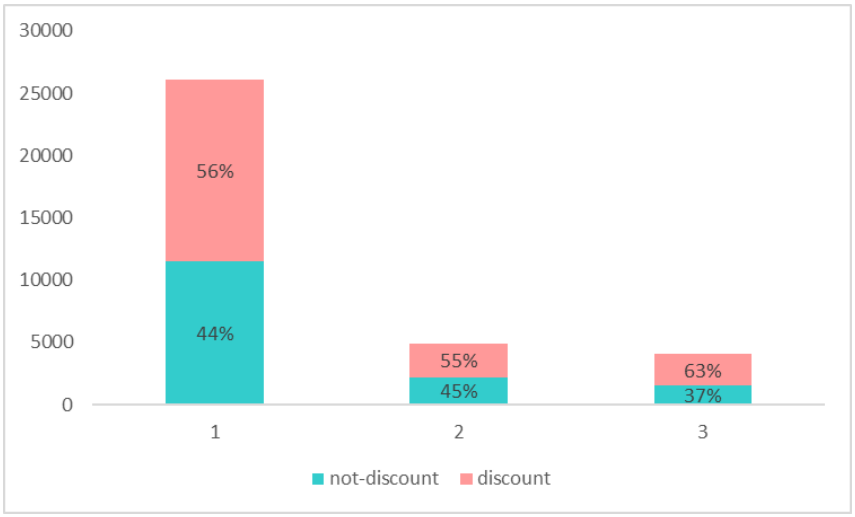

图 5-4 游戏支持系统与促销关系

在 Steam 游戏平台上，游戏支持的系统有 Windows 系统、Mac 系统和 Linux 系统，其中绝大多数游戏都支持在 Window 环境下运行，少量的游戏支持多系统环境运行。

从游戏支持的系统环境数量来看，支持一种系统的游戏数量最多。然而，通过图像不难发现支持三种系统的游戏在发行后一个月进行促销的比例是最高的，达到了 63%；相较于仅支持一种系统的促销率只有 56%。

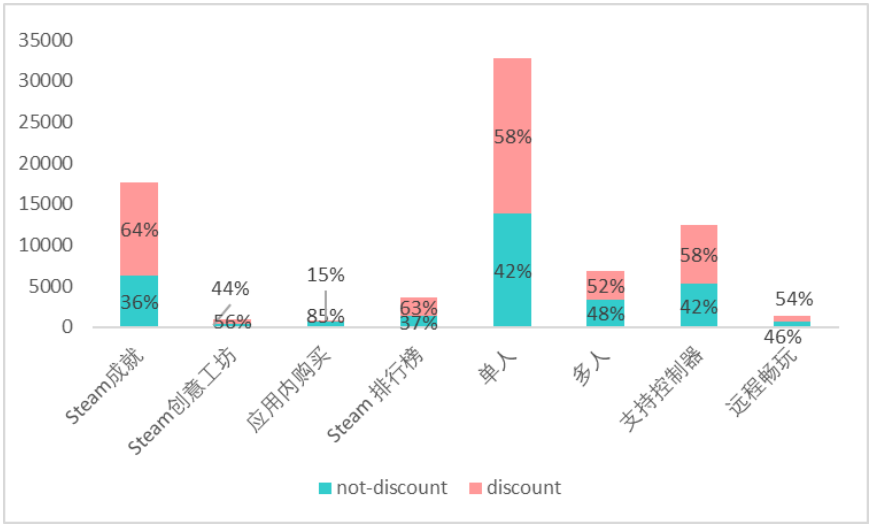

图 5-5 游戏特征与促销关系

从游戏具有的特征（features）来看，在该数据集中，单人游戏的数量非常多，有近 30000 个；支持应用内购买的游戏和拥有 Steam 创意工坊（支持玩家二次创作）的游戏数量相对较少，都分别仅有不到 5000 个。

从该图中可以发现，相较于单人游戏，多人游戏的促销比例更高，达到了 48%；而具有应用内购买特征的游戏在发行后一个月进行促销的概率最低，仅有 15%，考虑可能是由于具有该特征的游戏的受众往往已经有了在游戏市场的消费习惯，不会过多受促销的影响，所以游戏发行商对该类游戏的促销活动相对较慢。同时，可以发现处于 Steam 平台排行榜上的游戏有 63%都会在发行后一个月进行促销，猜测这些游戏可能是较大的游戏厂商发行的，质量和资金都相对有保障，所以会出现质量好、打折快的现象。另一方面，发现具有 Steam 成就的游戏发行一个月内促销率也比较高，达到了 64%，考虑可能是由于具有该特征的游戏和 Steam 平台的合作联系比较紧密，所以参加平台大促的次数多。

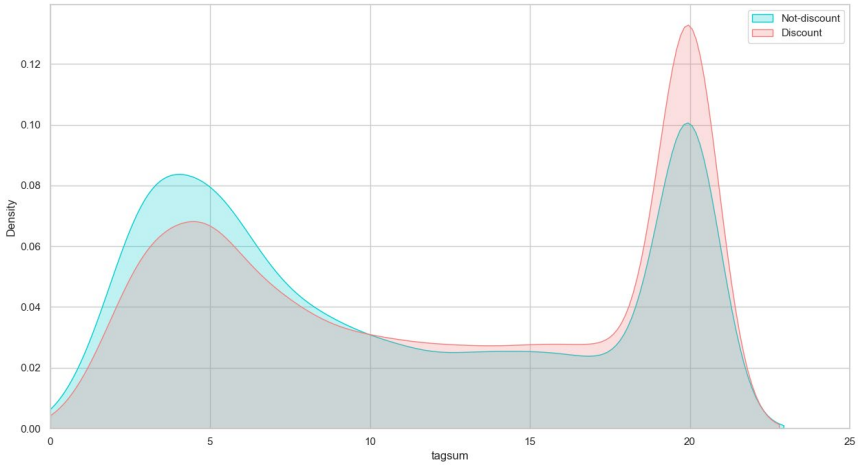

图 5-6 用户自定义标签数与促销关系

从用户自定义标签数来看，游戏数量出现了两个峰值，分别是在标签数为 4 左右时和标签数为 20 左右时。大多数普通游戏玩家只会给自己游玩过的游戏 3-4 个标签，因此在 4 左右出现峰值是合理的；而当一款游戏非常优秀、广受好评时，它就有可能具有更多用户因而拥有更多用户自定义标签。除这种情况外，应该是呈现用户自定义标签越多的游戏数量越少的趋势，也符合上图的表现。

在用户标签较少时，发行后一个月内不进行促销的游戏占比相对较高；而随着用户自定义标签数量的增多，促销的游戏数量也不断增多，最终超过不促销的游戏数量。由于用户自定义标签可以在一定程度上反映某款游戏的受众面程度，因此也可以推测越受众面越广的游戏打折概率越高。

2. 游戏的玩家信息

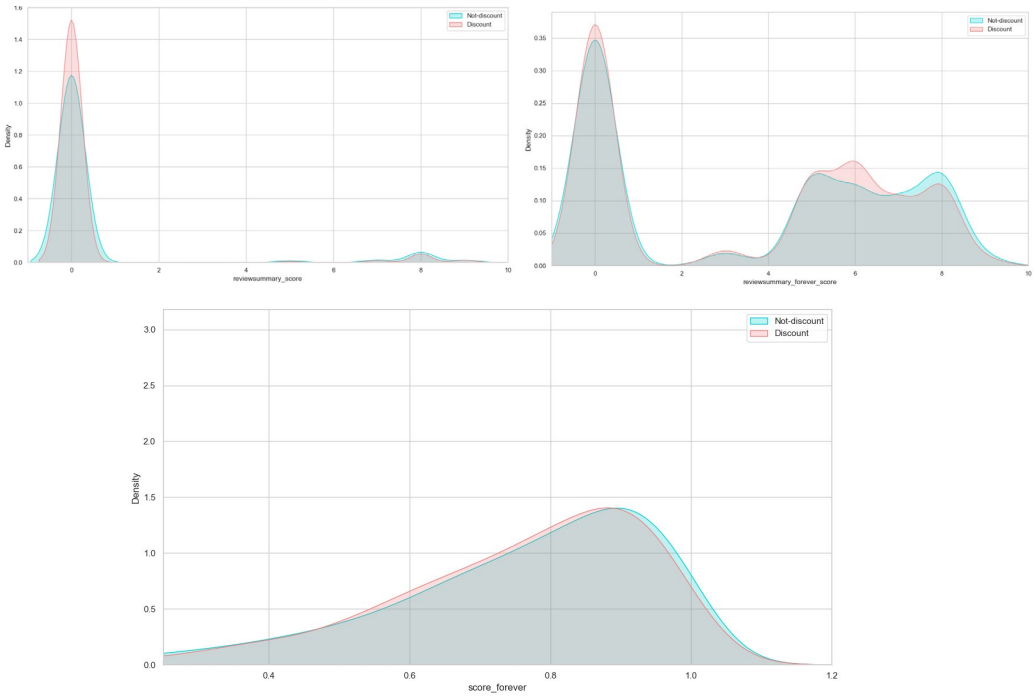

图 5-7 游戏好评率与促销关系

从玩家对游戏的好评率来看，可以发现无论是近 30 天内还是自游戏发售以来，好评率较低的游戏中打折游戏占比较大，而好评率较高的游戏中在发行一个月内不打折的较多。这与发行商常见的营销策略也非常符合，当客户不喜欢某款产品时，发行商倾向于尽快降低产品的价格来达到增加销量的目的；而当某款产品广受好评时，即便不进行促销也不会影响该款产品的销售情况了。

从玩家给游戏打分情况来看，游戏总体数量呈现左偏分布的特点。多数用户倾向于给游戏评分中上，但也有相当一部分用户给一些游戏打出了非常低的分数，这和我们上面看到的好评率的分布情况时一致的。而对于不同评分的游戏的促销情况来说，评分低的游戏倾向于在游戏发售后就进行促销；评分高的游戏则倾向于不促销或是推迟促销时间，再一次验证了

上文的推测。

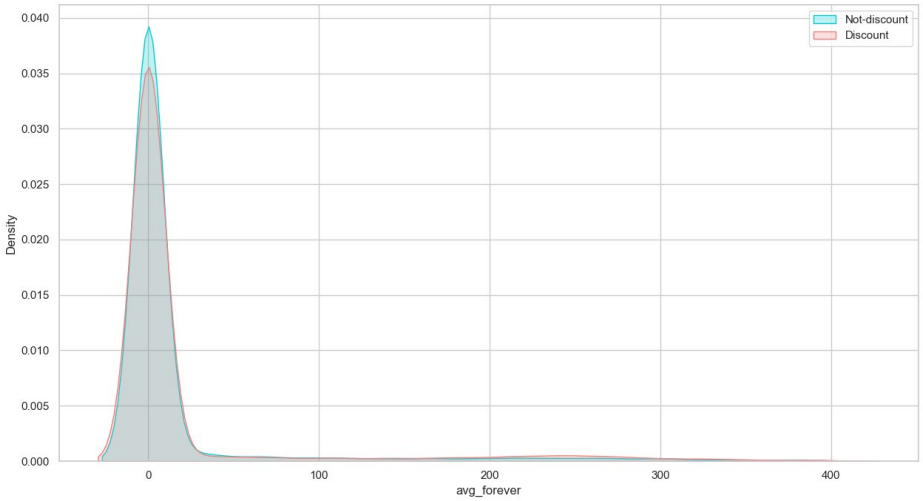

图 5-8 玩家游玩时长与促销关系

从玩家平均游玩时长来看，玩家平均游玩时长呈现明显的右偏趋势，说明绝大多数游戏是长期处于无人问津的状态的，而仅有极少数的优秀游戏能够吸引到玩家，使玩家花费大量的时间游玩。同时可以发现，平均游玩时长越短的游戏在短期内打折的比例越小，而平均游玩时长更长的游戏的发售首月打折比例则相差无几，甚至进行促销的比例还要略胜一筹。

六、模型构建

基于前述目标，本项目构建分类模型达到预测新发售的游戏是否会在发售首月进行促销活动的目标；针对每个模型进行训练，得到特征的重要性排序，达到探索促销游戏特征的目标。进一步，本项目构建回归模型达到更精准地预测游戏在发售后会经过多久才会打折到 50%以下。

本文所使用的模型有决策树模型、随机森林模型、GBDT 模型、xgboost 模型和多元线性回归模型。

1. 模型介绍

1）多元线性回归模型

多元线性回归模型的一般形式为：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \mu_i \; i = 1,2,\ldots,n$$

其中 k 为解释变量的数目，β_j（j=1,2,…,k)称为回归系数(regression coefficient)。上式也被称为总体回归函数的随机表达式。它的非随机表达式为

$$E(Y \mid X_{1i}, X_{2i}, \ldots X_{ki},) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

β_j也被称为偏回归系数（partial regression coefficient）

因为在判断游戏多久会进行促销的问题中，促销时间受游戏客观条件的影响，因此考虑使用多元线性模型，以促销时间距发售时间的天数做为因变量，其他因素做为自变量拟合模型，以期可以根据游戏客观条件对游戏促销时间进行估计。

2）**决策树模型**

本项目所使用的决策树模型为二叉分类树。二叉分类树采用基尼系数作为最优特征选择的度量标准。以分类系统为例，数据集 D 中类别 C 可能取值$c_1, c_2, \ldots, c_k$（k 是类别数），一个样本数据类别c_i的概率为p_i，那么概率分布的基尼指数公式表示为：

$$Gini(D) = 1 - \sum_{i=1}^{k} p_i^2$$

Gini(D)的物理含义：数据集 D 的不确定性，数值越大，表明其不确定性越大（与信息熵类似）。

如果 k=2，对应的基尼系数为$Gini(D) = 2p(1-p)$

如果数据集 D 根据特征 f 是否取某一可能值 f*，将 D 划分为D_1, D_2两部分，则：

$$D_1 = \{(x,y) \in D | f(x) = f^*\}, D_2 = D - D_1$$

那么特征 f 在数据集 D 上的基尼指数定义为：

$$Gini(D, f = f^*) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

在实际操作中，通过遍历所有特征（如果是连续的，需作离散化）及其取值，选择基尼指数最小所对应的特征和特征值。

3）**随机森林模型**

随机森林(Random Forest，RF）是 Bagging 算法的一种。Bagging 也叫自举汇聚法（bootstrap aggregating），是一种在原始数据集上通过有放回抽样重新选出 k 个新数据集来训练分类器的集成技术。它使用训练出来的分类器的集合来对新样本进行分类，然后用多数投票或者对输出求均值的方法统计所有分类器的分类结果，结果最高的类别即为最终标签。此类算法可以有效降低偏差，并能够降低方差。

随机森林是由很多决策树构成的并行集成算法，不同决策树之间没有关联。

当进行分类任务时，新的输入样本进入，就让森林中的每一棵决策树分别进行判断和分类，每个决策树会得到一个自己的分类结果，随机森林会用决策树的结果进行投票，票数最多的分类结果就成为最终随机森林的结果。

随机森林作为一种在机器学习中常用的算法有以下几种特点：

随机森林使用了 CART 决策树作为弱学习器。换句话说，其实我们只是将使用 CART 决策树作为弱学习器的 Bagging 方法称为随机森林。

同时，在生成每棵树的时候，每个树选取的特征都仅仅是随机选出的少数特征，一般默认取特征总数的开方。而一般的 CART 树则是会选取全部的特征进行建模。因此，不但特征

是随机的，也保证了特征随机性。随机森林采用的 CART 决策树是基于基尼系数选择特征的，基尼系数的选择的标准就是每个子节点达到最高的纯度，即落在子节点中的所有观察都属于同一个分类，此时基尼系数最小，纯度最高，不确定度最小。

由于随机性，对于降低模型的方差很有作用，故随机森林一般不需要额外做剪枝，即可以取得较好的泛化能力和抗过拟合能力。在样本量较大的情况大，随机森林的方法对于拟合和预测效果较好，对于分类响应变量的影响因素解读更为简洁、直观，易被理解。

4）GBDT 模型

GBDT，全称为 Gradient Boosting Decision Tree，也称作梯度提升决策树，和随机森林一样是一种基于决策树的集成学习算法。和大多数 Boosting 模型一样，GBDT 也是一种加法模型，它会串行地训练一组 CART 回归树，然后对所有树的结果加权取和获得最终结果，其中每一棵树都会拟合当前损失函数的负梯度方向。虽然 GBDT 中所有的树都是回归树，但经过合适的调整，它也可以被用来解决分类问题。

如果用 M 表示决策树的数量。$\mathrm{f_m(x_i)}$表示第 m 轮训练之后的整体，$\mathrm{f_m(x_i)}$即为最终输出的 GBDT 模型。第一步，创建第一棵树$\mathrm{f_1}(x)$，它是直接用回归树拟合目标值的结果，得到：

$$\mathrm{f_1}(x) = argmin \sum_{i=1}^{N} L(y_i, c)$$

对于第 2 到第 m 棵回归树，要计算出每一棵树的前面结果的残差：

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

对于当前的第 m 棵子树而言，需要遍历它的可行的切分点以及阈值，找到最优的预测值 c 对应的参数，使得尽可能逼近残差。最终得到回归树。

5）XGBoost 模型

XGBoost 也是梯度提升树模型的一种，同样是串行地生成模型，取所有模型的和为输出。XGBoost 将损失函数作二阶泰勒展开，利用损失函数的二阶导数信息优化损失函数，根据损失函数是否减小来贪心的选择是否分裂节点。同时，XGBoost 在防止过拟合方面加入了正则化、学习率、列采样、近似最优分割点等手段。在处理缺失值方面也做了一定的优化。

XGBoost 虽然和 GBDT 类似，也是一个加法模型，但是在每个结点分裂的标准上有所不同。XGBoost 对于当前节点计算出分裂前损失，再计算按照某一特征产生分裂之后的分裂后损失——即分裂后两个子树的损失和，并且定义分裂增益=分裂前损失-分裂后损失，如果分裂增益大于零，就可以认为分裂是可行的。

2. 决策树

从原有数据集的观测中抽取 80%作为训练集拟合模型，用剩余的 20%作为测试集。构建 CART 决策树模型，以基尼指数作为最优特征选择的度量标准，直接使用交叉网格搜索来优化决策树模型，边训练边优化，进行 4 折交叉检验。

构建如下决策树：

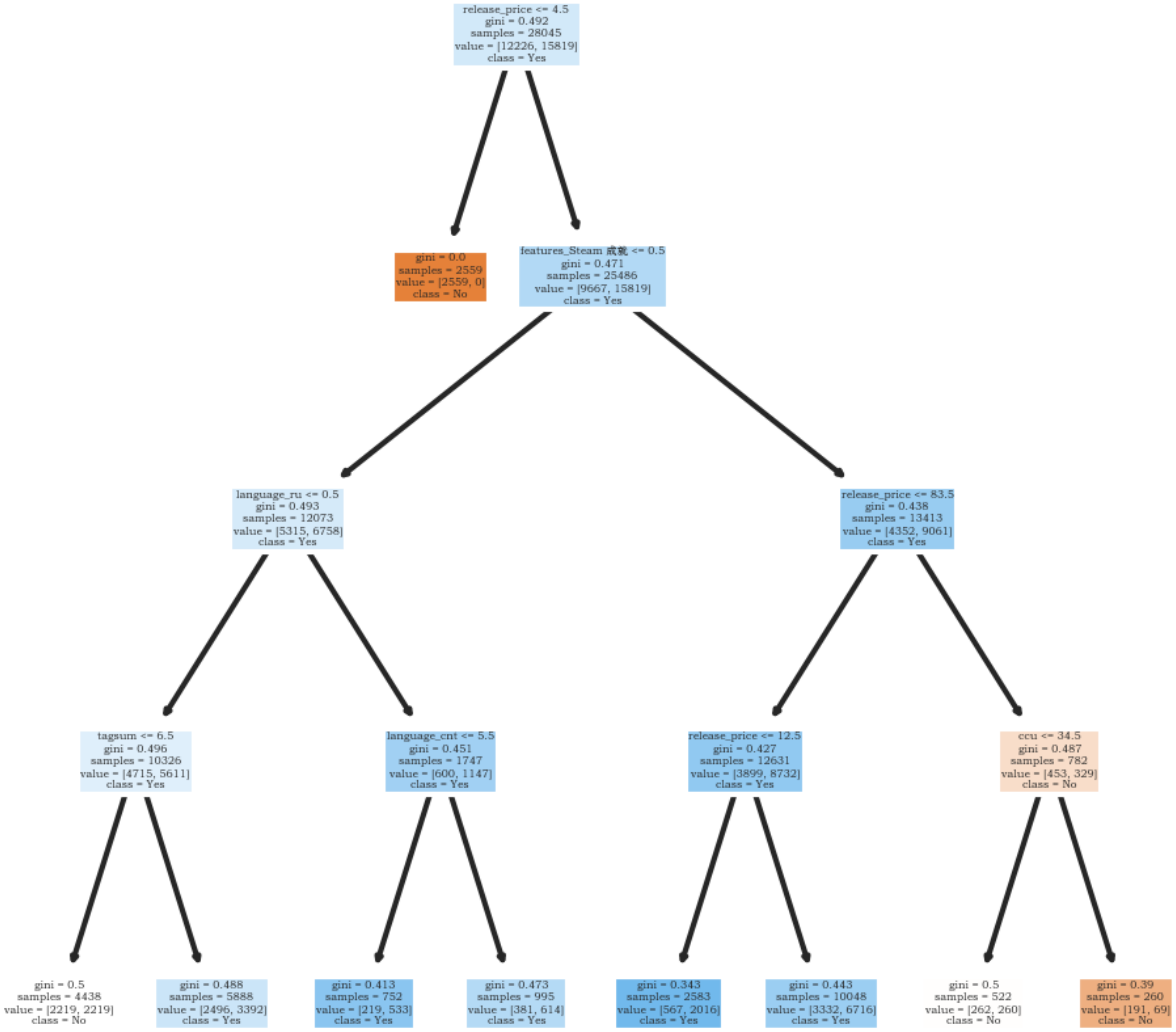

图 6-1 决策树结果

从叶子节点可以看出，游戏发售价格大于 4.5 元、不在 Steam 平台上拥有成就功能、不支持俄语并且用户自定义标签数量小于 6.5 的游戏在发售后首月进行打折的概率是最小的；而发售价格大于 4.5 元小于 12.5 元并且在 Steam 平台上拥有成就功能的游戏在发售后首月进行促销活动的概率最大。同时，也可以发现根据上述构建的决策树模型，如果发售价格大于 83.5 元，则无论子叶子节点同时在线人数峰值（ccu）的判断为何，该游戏都不会在发售后首月进行促销活动。根据上述模型可以发现游戏的发售价格是影响游戏在发售后是否首月促销的很重要的因素之一。

然而，决策树的预测精度通常不够好。因为它的变化幅度比较大，对训练数据的不同拆分方式可能会生成截然不同的决策树模型。在上面构建的决策树模型中，模型的 AUC 值只有 0.682。

3. 随机森林

为了进一步增大模型的平稳性，解决决策树高方差的问题。接下来建立随机森林模型。

在随机森林模型中，将进一步对相关参数进行调整，以提高整个模型的性能并避免模型过于复杂出现过拟合的情况。

随机森林主要的参数有子树的数量、树的最大生长深度、叶子的最小样本数量、分支节点的最小样本量、最大选择特征数。其中，子树的数量增加可以在一定程度上提高模型的性能，提高该参数不会影响单颗树的性能，只会增加整个模型的平稳性。对树的最大生长深度进行合理限制可以避免模型过于复杂、避免出现过拟合现象。如果不对树的深度进行限制，模型的树的深度可能会过大，一般来说 10~20 层的树深已经是比较复杂的模型了。同时，因为随机森林在构建每一颗树时并不会用到所有的特征变量，因此找到最合适的最大选择特征数也是非常重要的。我们使用画学习曲线和网格搜索的方法来寻找最优参数

由上文可以知道，决策树和随机森林模型在建立模型的时候并不会用到所有的样本数据，而模型的表现好坏与初始数据的划分结果由很大的关系，为了处理这种情况，我们采用交叉验证的方式来减少偶然性。

因为树的个数是对整个随机森林影响程度最大的参数，先对其进行调整。每隔十步建立一个随机森林，获得在不同的 n_estimator 情况下模型的得分。画出学习曲线如下图所示。当 n_estimator 从 0 开始增大到 15 时，模型准确度有肉眼可见的提升；而当子树的数量越来越大时，准确率发生波动，当取值为 81 时，模型获得最大得分为 0.675。

接下来将子树数量参数的调整范围缩小到 81 左右，并把步长由 10 缩小到 1，以期获得更好的取值。画出的学习曲线如下图所示。可以发现模型得分在 81 附近非常波动，但仍然是在 81 的时候取到最大值，为 0.675.

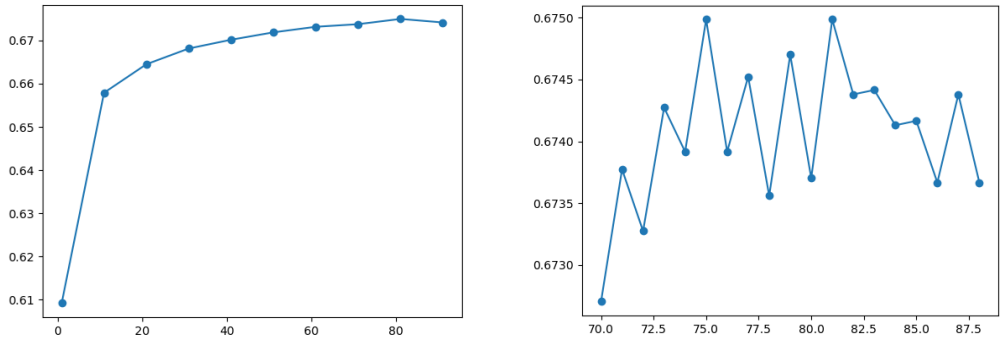

图 6-2 随机森林 n_estimator 学习曲线

接下来的调参方向是使模型复杂度减小的方向，从而接近泛化误差最低点。我们使用能使模型复杂度减小，并且影响程度其次的树的最大深度（max_depth）。直接使用网格搜索来确定最大深度的最优取值，可以得到结果最佳深度为 18 的时候，最大得分为 0.679，进一步提高了模型的准确率。

之后同样使用网格搜索法对剩余的参数一一进行调整，最终获取最佳的最大特征数量占比为 0.4，叶子的最小拆分样本量为 20，此时模型准确率进一步优化到 0.687。AUC 得分为 0.734。

根据随机森林模型结果可以对特征重要性基于基尼系数进行排序，特征重要性图如下图所示，其中重要性前十的特征的数据如下表所示：

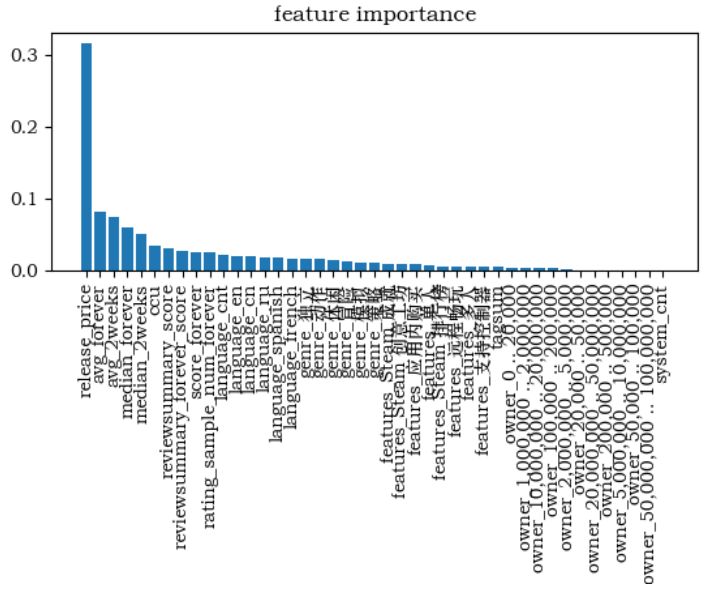

图 6-3 随机森林特征重要性

表格 6-1 随机森林特征重要性数值

release_price	0.315
avg_forever	0.081
avg_2weeks	0.075
median_forever	0.061
median_2weeks	0.05
ccu	0.034
reviewsummary_score	0.031
reviewsummary_forever_score	0.027
score_forever	0.026
rating_sample_num_forever	0.025

从图表中可以看出，对于游戏发售首月是否打折影响较大的变量为发售价格、平均游玩时长、两周内平均游玩时长、游玩时间中位数、同时在线峰值人数、游戏好评率、游戏得分和游戏评分人数。其中，发售价格对游戏发售首月是否促销的影响程度明显大于其他变量，这说明在购买游戏的时候可以重点考虑游戏定价、游戏受欢迎程度和游戏评分的影响。

4.GBDT 模型

GBDT 和随机森林一样都是一种集成算法。然而，随机森林采用的是 bagging 思想而 GBDT 采用的是 boosting 思想，即随机森林采用的是有放回的均匀抽样，而 GBDT 采用的是根据错误率来抽样。为了探明哪种模型在本数据集上效果最好，接下来建立 GBDT 模型。

使用所有数据的 80%作为训练集，20%作为测试集。首先在不预先设定任何参数的情况下建立模型，得到模型的准确率为 0.67，AUC 为 0.70。GBDT 中较为重要的参数有 n_estimators，max_depth，min_sample_split，min_samples_leaf，通过网格搜索的方法依次对这四个参数搜寻最优值。最终设定参数分别为 learning_rate=0.1，n_estimators=70，max_depth=16，min_sample_split=3，min_samples_leaf=10。得到模型的 AUC 为 0.7222。

根据 GBDT 模型的结果可以对特征重要性进行排序，结果如下图所示：

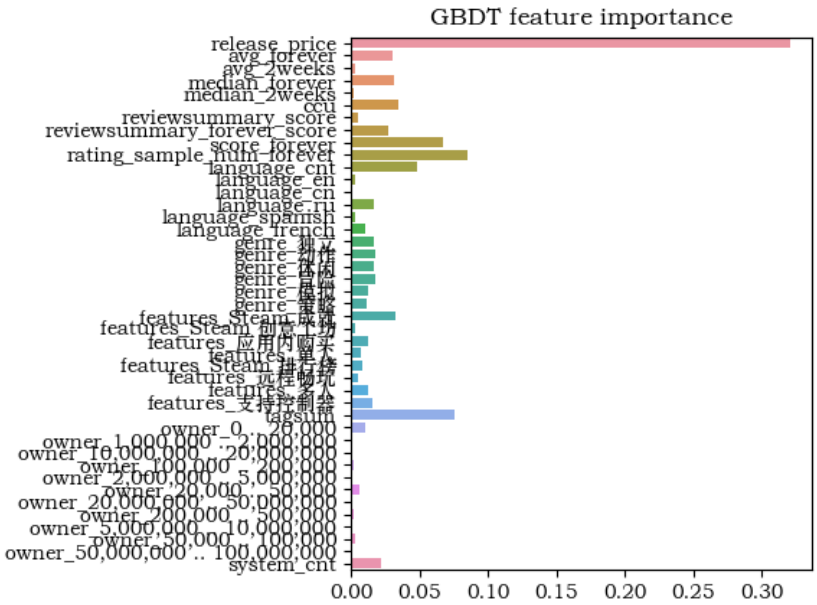

图 6-4 GBDT 模型特征重要性排序图

从图中可以看出，对游戏发行第一个月是否会进行促销影响较大的特征排序前五的依次是：发行价格（release_price），总评分人数（rating_sample_num_forever），用户自定义标签数（tagsum），用户总评分（score_forever），最大同时游玩人数（ccu）。不难发现这五个特征均和用户行为相关，GBDT 模型的结果表明，游戏是否会在发行首月进行促销主要取决于玩家反馈，即发行商更可能通过发行后玩家的总数、评分、活跃程度等因素制定促销策略。

5. XGBoost 模型

虽然 XGBoost 和 GBDT 同属于 boosting 类的树模型，某种意义上 XGBoost 可以看作是 GBDT 的一种变体，但 XGBoost 模型在结点分类标准上和 GBDT 有所不同。相较于 GBDT，XGBoost 对代价函数中同时用到了一阶和二阶函数，并且在代价函数里加入了正则项。因此 XGBoost 模型有可能会优于 GBDT 模型，为了探明哪种模型在本数据集上效果最好，接下来建立 GBDT 模型。

使用所有数据的 80%作为训练集，20%作为测试集。XGBoost 中较为重要的参数有 n_estimators，max_depth，learning_rate，通过网格搜索的方法依次对这三个参数搜寻最

优值。最终设定参数分别为 learning_rate=0.05, n_estimators=200, max_depth=15,。得到模型的 AUC 为 0.6881。

根据 XGBoost 模型的结果可以对特征重要性进行排序，结果如下图所示：

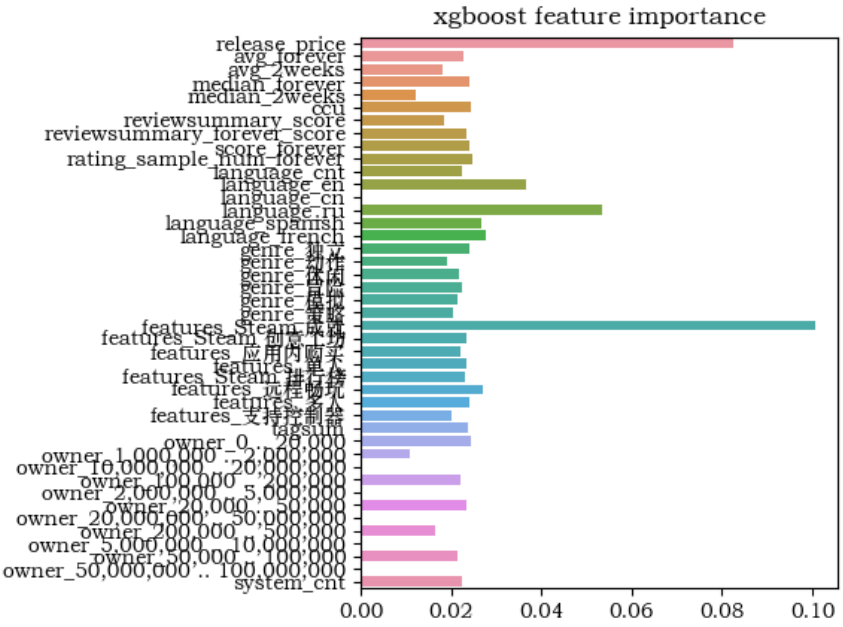

图 6-5 xgboost 模型特征重要性排序图

从图中可以看出，对游戏发行第一个月是否会进行促销影响较大的特征排序前五的依次是：steam 成就（features_steam 成就），发行价格（release_price），是否支持俄语（language_ru），是否支持英语（language_en），是否支持远程畅玩（features_远程畅玩）。，XGBoost 模型的结果表明，游戏是否会在发行首月进行促销主要取决于游戏本身的丰富性和与 Steam 平台的合作情况。

6. 多元线性回归

进一步，由于通常新发售的游戏的第一次促销的力度非常有限，我们还希望探究一个游戏在发售后多久可以打折超过 50%。为了探究半价促销时间和游戏的各种客观因素的关系，本文首先考虑使用线性回归模型来探究该问题。

在所有数据中，仅有 18395 个游戏的最低折扣达到过 50%，将没有促销超过 50%的游戏数据条目删除。下面的线性模型在 18395 条数据上进行。

首先，为了探究因变量（新发售游戏半价促销所需要的时间）和各个自变量之间是否存在线性关系，本文分别检验了连续和分类自变量与因变量之间的相关关系。对于连续型自变量，采取计算 Person 相关系数的方法，结果如下图所示：

表格 6-2 连续变量和因变量的相关系数表

	release_pri	avg_foreve	avg_2week	median_fo	median_2v	ccu	reviewsum	reviewsum	score_forev	rating_sam	language_c	tagsum	first_discou	fifty_discou
release_price	1	0.190611	0.095055	0.105738	0.095938	0.009758	0.306605	0.232658	0.195927	0.019212	0.146825	0.159315	0.039261	0.327584
avg_forever	0.190611	1	0.201816	0.881905	0.189461	0.288506	0.219615	0.117283	0.097526	0.29198	0.081226	0.075773	0.019811	0.050256
avg_2weeks	0.095055	0.201816	1	0.118829	0.990224	0.079166	0.18717	0.062993	0.051762	0.083185	0.049207	0.049069	0.007689	0.033516
median_forever	0.105738	0.881905	0.118829	1	0.117797	0.046211	0.120706	0.076895	0.063836	0.048345	0.038303	0.045427	0.005568	0.029543
median_2weeks	0.095938	0.189461	0.990224	0.117797	1	0.042376	0.191491	0.064717	0.053215	0.045118	0.04801	0.050844	0.008309	0.037381
ccu	0.009758	0.288506	0.079166	0.046211	0.042376	1	0.05443	0.017278	0.014481	0.993702	0.044396	0.01698	0.014511	0.005794
reviewsummary_scor	0.306605	0.219615	0.18717	0.120706	0.191491	0.05443	1	0.310397	0.260741	0.080414	0.202666	0.246558	0.045238	0.539326
reviewsummary_forev	0.232658	0.117283	0.062993	0.076895	0.064717	0.017278	0.310397	1	0.968292	0.026788	0.171141	0.231502	0.019331	0.543845
score_forever	0.195927	0.097526	0.051762	0.063836	0.053215	0.014481	0.260741	0.968292	1	0.022631	0.157085	0.219037	0.015292	0.192543
rating_sample_num_	0.019212	0.29198	0.083185	0.048345	0.045118	0.993702	0.080414	0.026788	0.022631	1	0.052463	0.025535	0.01669	0.010742
language_cnt	0.146825	0.081226	0.049207	0.038303	0.04801	0.044396	0.202666	0.171141	0.157085	0.052463	1	0.19253	-0.00808	0.02563
tagsum	0.159315	0.075773	0.049069	0.045427	0.050844	0.01698	0.246558	0.231502	0.219037	0.025535	0.19253	1	-0.04518	0.008759
first_discount_period	0.039261	0.019811	0.007689	0.005568	0.008309	0.014511	0.045238	0.019331	0.015292	0.01669	-0.00808	-0.04518	1	0.665747
fifty_discount_period	0.327584	0.050256	0.033516	0.029543	0.037381	0.005794	0.539326	0.543845	0.192543	0.010742	0.02563	0.008759	0.665747	1

可以发现只有 reviewsummary_score, reviewsummary_forever_score, release_price, first_discount_period 和因变量 fifty_discount_period（半价促销距发售时长）有较为明显的线性关系。为了提高模型的拟合优度和预测准确性，将其他没有明显线性关系的变量从模型中剔除。

对于分类自变量，采取方差分析的方法判断它们和因变量之间的相关关系，结果如下表所示：

表格 6-3 分类变量方差分析结果

变量名称	F	PR(>F)
C(language_en)	0.471191	0.492448
C(language_ru)	0.00058	0.980787
C(language_spanish)	5.297127	0.021372
C(language_french)	158.9675	2.69E-36
C(genre_独立)	22.25616	2.40E-06
C(genre_动作)	27.64153	1.48E-07
C(genre_休闲)	92.4134	7.92E-22
C(genre_冒险)	3.832846	0.050273
C(genre_模拟)	21.12632	4.33E-06
C(genre_策略)	0.873547	0.349987
C(features_steam_achievement)	80.5348	3.13E-19
C(features_steam_workshop)	52.71421	4.01E-13
C(features_应用内购买）	0.10965	0.740547
C(features_单人)	6.914778	0.008556
C(features_steam_rank)	0.617611	0.431947
C(features_远程畅玩)	84.99422	3.31E-20
C(features_多人)	12.97449	0.000317
C(features_支持控制器)	40.41345	2.10E-10

从上表可以发现在 5%的显著性水平下，支持英文（language_en）、支持俄语（language_ru）、冒险类别（genre_冒险）、策略类别（genre_策略）、应用内购买特征

（features_应用内购买）、Steam 排行榜特征（features_steam_rank）是不显著的，可以拒绝原假设，认为它们和因变量（游戏发售后半价促销的时间）是不具有显著的线性关系的。将这些变量从模型中剔除。

之后，以调整 R 方作为标准，用向前的逐步回归拟合多元线性回归模型并进一步进行变量选择。得到结果如下所示：

表格 6-4 逐步回归结果

变量名	coef	p
intercept	103.6	<0.001
first_discount_period	0.9	<0.001
reviewsummary_forever_score	13.6	<0.001
release_price	0.44	<0.001
first_month_discount	45.8	<0.001
reviewsummary_score	8.54	<0.001
language_ru	-37.1	<0.001
genre_动作	-29.3	<0.001
features_支持控制器	17	<0.001
features_steam_workshop	54	<0.001
genre_休闲	-22.1	<0.001
features_应用内购买	-96	<0.001
tagsum	-1.4	<0.001
language_french	22.4	<0.001
features_steam_achievement	16	<0.001
genre_模拟	15.7	0.001
features_远程畅玩	29.3	0.001
genre_冒险	-10.1	0.008
language_spanish	-24.2	0.010
language_en	37.3	0.095
features_多人	7.8	0.116
genre_策略	7.5	0.118

可以发现以调整 R 方为标准的逐步回归剔除了 13 个变量，剩余的大多数变量在 5%的显著性水平下都是显著的，只有 fetures_多人、genre_策略、language_en 三个变量在 5%的显著水平下不显著。该模型的 F=336.6，F 检验对应的 p 值<0.01，说明该模型整体在 5%的显著性水平下也是显著的。模型的调整 R 方为 0.48，模型的拟合效果略低，但经过尝试对变量的多种变换也并没有能有效提高 R 方，可能是受限于数据本身。同时，经过计算 VIF，

发现该模型中所有变量的 VIF 都小于 10，说明不存在比较严重的多重共线性问题。

得到的线性模型的表达式如下所示：

$$
\begin{aligned}
\text{fifty\_discount\_period} = {} & 0.86\text{first\_discount\_period} + 13.6\text{reviewsummary\_forever\_score} \\
& + 0.44\text{release\_price} + 45.8\text{first\_month\_discount} \\
& + 8.5\text{reviewsummary\_score} - 37.1\text{language\_ru} - 29.3\text{genre\_动作} \\
& + 17.0\text{features\_支持控制器} + 54.0\ \text{features\_steam\_workshop} \\
& - 22.2\text{genre\_休闲} - 96.0\text{features\_应用内购买} - 1.4\text{tagsum} \\
& + 22.4\ \text{language\_french} + 16.0\text{features\_steam\_achievement} \\
& + 15.7\ \text{genre\_模拟} + 29.2\text{features\_远程畅玩} - 10.1\text{genre\_冒险} \\
& - 24.2\ \text{language\_spanish} + 7.8\text{features\_多人} + 7.5\text{genre\_策略} + 103.6
\end{aligned}
$$

该模型说明：

控制其他因素不变，第一次促销距发售日期的时间每增加一天，促销至半价的时间就增加 0.86 天。说明游戏的促销策略是一以贯之的，如果在发售最初促销较晚，那么后续的促销进程也会比较慢，打折力度也会相对较小。

控制其他因素不变，游戏好评率每增加 1，促销至半价的时间就增加 13.6。说明如果一个游戏广受好评，那么游戏厂商大概率不用担心销量问题，也就没有了加大促销力度和提快促销进程的动力。相反，如果一个游戏的玩家好评率较低，那么厂商很有可能为了提高销量而加快将游戏打折至发售价格的 50%。

控制其他因素不变，在 Steam 平台拥有创意工坊的游戏比没有的促销至半价的时间要多 54 天。由于创意工坊的存在，游戏即便是没有厂商更新内容也会有源源不断的玩家自制内容，而这样的全新的游戏内容会让游戏即便在没有官方更新的情况下也保持新鲜感，自然不愁销量，也就没有必要早早通过较大的折扣来吸引玩家了。

控制其他因素不变，有应用内购买的游戏要比没有的促销至半价的时间要少 96 天。如果一个游戏有游戏内付费的机制，通常情况下玩家为这款游戏花的钱大多在游戏过程中而并非在最初购买游戏时。在这种情况下，开发商为了最大化利益大概率会选择降低游戏售价而提高游戏内购买的内容价格。

控制其他因素不变，支持远程畅玩的游戏要比不支持的促销至半价的时间多 29.2 天。通常支持远程畅玩（电视、平板、手机）的游戏因为要适配多种设备，其开发费用会相对较大，如果过早进行半价促销可能会使开发商无法收回成本。因此这类开发成本较高的游戏无论是促销进程不会太快。

从模型结果中也可以看出，对于不同类别的游戏（冒险、模拟、策略、休闲等），游戏类别对半价促销进程的影响也不甚一致。一方面这有可能与不同种类游戏的生命周期、平均游戏时长、制作成本等有关，另一方面不同种类的游戏往往受众群体也不完全一致，因此开

发商在进行促销策略的制定的时候必然会因地制宜。

七、模型评估和比较

1.模型评估

为了模型的进一步开发，同时检验模型在预测的中的效果，进行模型评估是必要的。本项目采用的模型评估方式主要是 ROC 曲线。

1. ROC 曲线介绍

ROC 的全称是 Receiver Operating Characteristic Curve，中文名字叫“受试者工作特征曲线”， 该曲线的横坐标为假阳性率（False Positive Rate，FPR）

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{N}$$

其中 FP 为实际为假而被预测为真的观测数量，TN 是实际为假并且被预测为假的观测数量，N 是实际为假的总数量。

纵坐标为真阳性率（True Positive Rate，TPR）

$$TPR = \frac{TP}{TP + FN} = \frac{TP}{P}$$

其中 TP 是实际为真且被预测也为真的观测数量，FN 是实际为真而被预测为假的观测数量，P 是实际为真的总数量。

针对模型预测结果的每一个概率值，可以计算该概率值下的混淆矩阵，从而得到每一个观测下的 FRP 和 TPR，将得到的数据绘制成散点图并连上折线即得到 ROC 曲线。

ROC 曲线与横轴以及 y=1 这一条轴所围成的面积称为 AUC，这是衡量模型拟合效果的一个重要指标。

2. 随机森林模型的 ROC 曲线

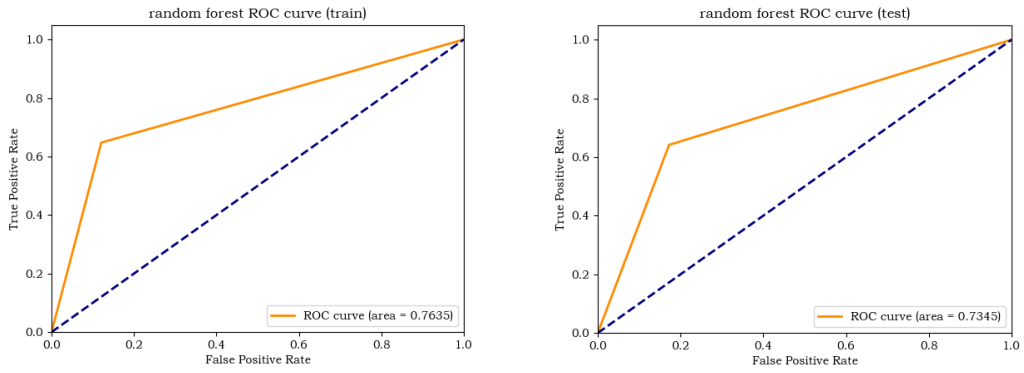

图 7-1 随机森林训练集和测试集 ROC 曲线

从训练集和测试集的 ROC 曲线可以看出，随机森林模型在训练集上的 AUC 为 0.7635，在测试集上的 AUC 为 0.7345，两者数值相对较高且相差不大，说明随机森林具有较高的拟

合效果，且不存在过拟合现象。

3. GBDT 的 ROC 曲线

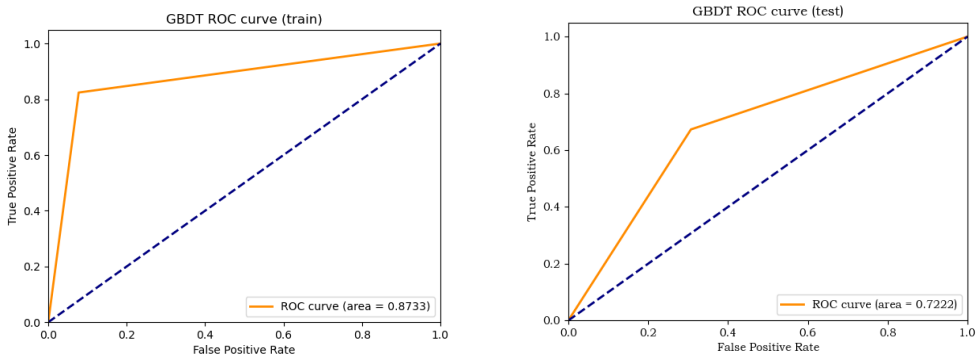

图 7-2 GBDT 训练集和测试集的 ROC 曲线

从训练集和测试集的 ROC 曲线可以看出，GBDT 模型在训练集上的 AUC 为 0.8733，在测试集上的 AUC 为 0.7222。该模型在训练集上的 AUC 明显大于在测试集上的 AUC，说明 GBDT 模型在训练集上存在一定程度的过拟合现象。

4. XGBoost 的 ROC 曲线

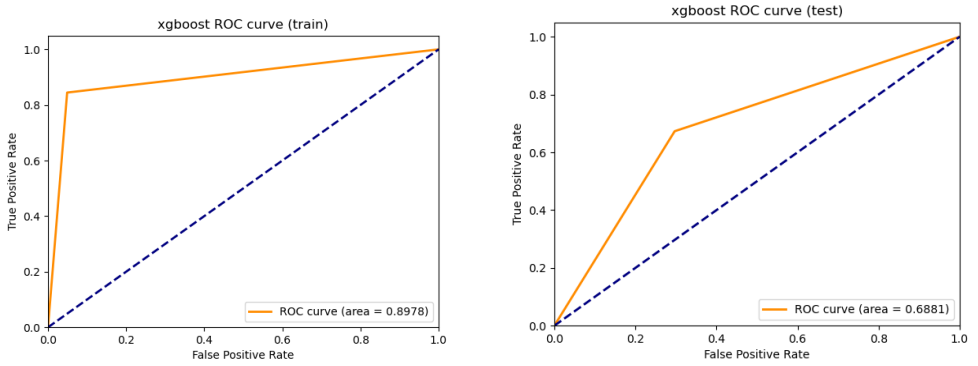

图 7-3 xgboost 训练集和测试集的 ROC 曲线

从训练集和测试集的 ROC 曲线可以看出，GBDT 模型在训练集上的 AUC 为 0.8978，在测试集上的 AUC 为 0.6881。该模型在训练集上的 AUC 明显大于在测试集上的 AUC，说明 GBDT 模型在训练集上存在过拟合现象。

2.模型比较与选择：

为了选择最优的模型，以达到最佳预测新游戏发行首月是否打折的目的，针对训练集准确率、测试集准确率、AUC 对四种模型进行比较，得到汇总结果表如下：

表格 7-1 模型准确率和 AUC 对比

	决策树	随机森林	GBDT	XGBoost
训练集准确率	0.6775	0.6870	0.8584	0.8821
测试集准确率	0.6634	0.6734	0.6788	0.6815
AUC	0.6823	0.7345	0.7222	0.6881

比较模型在测试集和训练集上的表现，决策树和随机森林模型的泛化能力都相对较强，但决策树模型的预测准确率略低于随机森林、AUC 明显低于随机森林。GBDT 和 XGBoost 的模型预测准确率虽然有了进一步提高，但是其泛化能力明显削弱，并且 AUC 值低于随机森林。

如果从基于准确率的学习曲线进行比较，通过五折交叉检验可以得到随机森林、GBDT 和 XGBoost 三种模型的学习曲线如下所示（横坐标为样本量，纵坐标为准确率）：

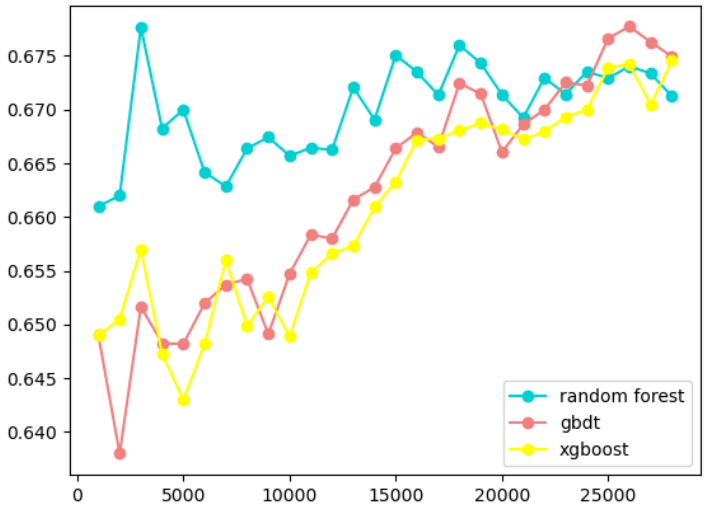

图 7-4 随机森林、GBDT、XGBoost 的学习曲线

通过上图可以发现在本电子游戏数据集上，样本量较小时随机森林的预测效果较好，且随着样本量增加预测准确率有轻微上升的趋势。GBDT 和 xgboost 模型虽然预测精度随样本量增加有明显的上升趋势，但在样本量较小（小于 20000）时预测精度明显低于随机森林。当样本量较大时三者的预测精度趋近。

综上所述，随机森林模型的泛化能力较强，AUC 也相对较高，并且在小样本和大样本下预测精度的表现都不错。在该问题下，我们倾向选择随机森林来进行预测并获得对游戏发行首月是否打折的重要影响因素。

八、结论与建议

通过上述模型分析，相较于决策树，可以采用稳定性更强的随机森林对新游戏发售首月是否打折进行预测，并可用多元线性模型对已发售游戏何时促销至半价进行较为精准的预测。根据模型结果，对消费者提出如下有关游戏促销的建议：

1）关注游戏的发行价格

在一般情况下，游戏的发行价格越高，其总体促销进程和促销至半价的速度就会越快。然而当考虑到游戏是否在第一个月内就会进行降价促销时，发行价格过高的游戏反而进行首月促销的概率较低。所以当消费者试图购买单价过高的产品时，如果想以一个较为优惠的价

格购入，就要做好无法“抢先体验”的准备。相反，对于一个发行价格不高的游戏来说，建议消费者在第一个月内找准促销机会就进行购入，因为定价低的游戏的首促会开展的相对较快，但长期来看也很难有较大力度的折扣幅度。

2）关注游戏的用户评价

游戏的用户评价是用户和开发商沟通的重要渠道，有些开发商会甚至会根据评价来选择性地对游戏进行更新和修改。用户评价数量少、评分低的游戏倾向于在发售首月进行促销；长期来看，这类游戏如果在长期的用户口碑没有明显的提升，其促销折扣力度也会非常大。因此，如果玩家心仪的游戏不被大多数人所看好，或许这就是用户入手该游戏的最好时机。

3）关注游戏内容的丰富度和自更新性

游戏的促销策略归根结底还是一种吸引新玩家的手段，如果游戏本身内容丰富或是拥有Steam 创意工坊这类允许所有玩家进行二次创作的功能，那么这款游戏长期看来促销折扣的力度不会太大，建议玩家趁早入手以享有更多的游戏内容的体验。

九、参考文献

[1] 瞿珊. 基于机器学习的网络游戏收益预测实证研究[D].重庆大学,2019

[2] 马明浩. 大数据时代用户游戏内付费预测研究[D].东华大学,2019

[3] 安俊峰. 游戏评价数据的分类预测研究[D].东华大学,2014.

[4] 汪洨. H 手机游戏公司某产品绩效前因要素的实证研究[D].对外经济贸易大学,2020.

[5] 肖磊. 中国网络游戏厂商营销策略研究[D].中国海洋大学,2013.

[6] 任静. 手机游戏精细化营销策略研究[D].北京邮电大学,2011.

[7] Nik Davis. Gathering Data from the Steam Store API using Python [EB/OL]. https://nik-davis.github.io/posts/2019/steam-data-collection/ 2019.05

[8] Le Roy Frédéric. Vertical vs horizontal coopetition and the market performance of product innovation: An empirical study of the video game industry[J]. Technovation. Volume 112, 2022.

声　　明

本人郑重声明所呈交的论文是我个人在指导老师的指导下进行的研究工作及取得的研究成果，不存在任何剽窃、抄袭他人学术成果的现象。我同意（ √ ）/不同意（　）本论文作为学校的信息资料使用。

论文作者（签名）　李晨禹

2022 年 5 月 17 日