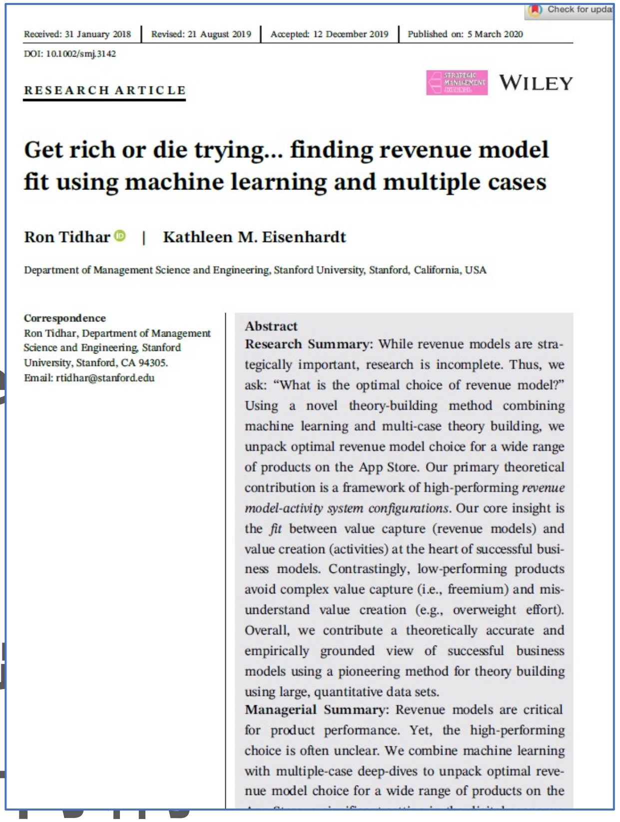
基于机器学习的商业模式研究

- 问题：如何设计合适的**收益模式(revenue models)**？
- 已有研究举出了三种收益模式：paid、advertising、free
- 方法概述：
  1. 探索性数据分析：揭露了数据趋势，并为接下来的案例采样提供了
  2. 多案例理论构建：识别理论构建与机制，并为接下来的机器学习埋
  3. 机器学习：基于大量数据的分析，进一步确定案例研究结果，并且扩充了发现



Received: 31 January 2018 | Revised: 21 August 2019 | Accepted: 12 December 2019 | Published on: 5 March 2020
DOI: 10.1002/smj.3142

RESEARCH ARTICLE

WILEY

Get rich or die trying... finding revenue model fit using machine learning and multiple cases

Ron Tidhar | Kathleen M. Eisenhardt

Department of Management Science and Engineering, Stanford University, Stanford, California, USA

Correspondence
Ron Tidhar, Department of Management Science and Engineering, Stanford University, Stanford, CA 94305.
Email: rtidhar@stanford.edu

Abstract
Research Summary: While revenue models are strategically important, research is incomplete. Thus, we ask: "What is the optimal choice of revenue model?" Using a novel theory-building method combining machine learning and multi-case theory building, we unpack optimal revenue model choice for a wide range of products on the App Store. Our primary theoretical contribution is a framework of high-performing *revenue model-activity system configurations*. Our core insight is the *fit* between value capture (revenue models) and value creation (activities) at the heart of successful business models. Contrastingly, low-performing products avoid complex value capture (i.e., freemium) and misunderstand value creation (e.g., overweight effort). Overall, we contribute a theoretically accurate and empirically grounded view of successful business models using a pioneering method for theory building using large, quantitative data sets.
Managerial Summary: Revenue models are critical for product performance. Yet, the high-performing choice is often unclear. We combine machine learning with multiple-case deep-dives to unpack optimal revenue model choice for a wide range of products on the

Tidhar, R., and Eisenhardt, K. M. 2020. "Get Rich or Die Trying... Finding Revenue Model Fit Using Machine Learning and Multiple Cases," *Strategic Management Journal* (41:7), pp. 1245-1273.

数据描述

App Store

苹果App Store

抽样

2015年11月的所有APP

受欢迎的APP

（13195个）

不受欢迎的APP

（53457个）

整理

APP名称	市场类别	欢迎度	收益模式
APP-1	金融	是	paid
APP-2	旅游	否	advertising
APP-3	游戏	是	freemium

市场类别：App Store自带的22种市场类别，如体育、旅行、生活等

欢迎度：如果某个APP在苹果App Store实时更新的”popular”名单中出现过，则这个APP是受欢迎的。

收益模式：把每个APP划分为**paid、advertising、freemium**中的一种，这可以根据App Store的”in-app purchase”得出

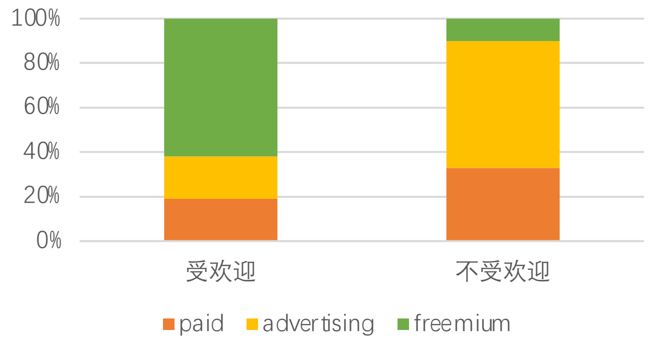
方法概述—探索性数据分析

APP名称	市场类别	欢迎度	收益模式
APP-1	金融	是	paid
APP-2	旅游	否	advertising
APP-3	游戏	是	freemium

频率统计

可视化

1、



2、各个市场类别下，三种收益模式占比

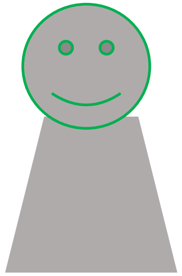
3、绘制freemium付费增值的频率图，发现其呈双峰分布。为后文将freemium划分为bundled freemium、fragmented freemium提供了思路

方法概述—多案例理论构建

APP名称	市场类别	欢迎度	收益模式
APP-1	金融	是	paid
APP-2	旅游	否	advertising
APP-3	游戏	是	freemium

1. 选取八大市场类别
2. 每一市场类别中抽取
    - 占比最大的收益模式中popular与unpopular各一个
    - 非占比最大的收益模式中popular一个
    - 知名的APP一个
    - 随机的一个unpopular APP

综上所述，共24个popular case，16个unpopular case



理论抽样

多源数据三角分析

案例内研究
跨案例研究

方法概述—多案例理论构建

结论与发现：

1. Advertising被扩充为third-party模式，因为广告商不是唯一的第三方付款人，有的APP会向餐厅收取顾客订餐费。
2. 用户资源(user resources)有价值时，适合third-party模式

   产品复杂性(product complexity)越高，越适合freemium模式

   线下品牌(offline brand)、营销(marketing)、设计(design)更适合paid模式

结论2为下一步的机器学习埋下了种子

方法概述—机器学习

- 在案例研究结论2中我们发现了user resources、offline brand、marketing、product complexity、 design等和收益模式密切相关的因素

- 能否用机器学习的方法，用这些因素“预测”收益模式。即：给出一个popular产品的user resources、offline brand、marketing、product complexity、 design等特征，能否通过机器学习的方法得知这个产品的收益模式

- 如果能，那么这些特征是如何影响“预测”的。即这些特征对收益模式的重要性程度排序？

- 这里选取三种机器学习算法：Penalized multinomial logistic regression、Decision tree、Random forest ensembles。这三种算法都能预测收益模式，且给出特征重要性排序

方法概述—机器学习

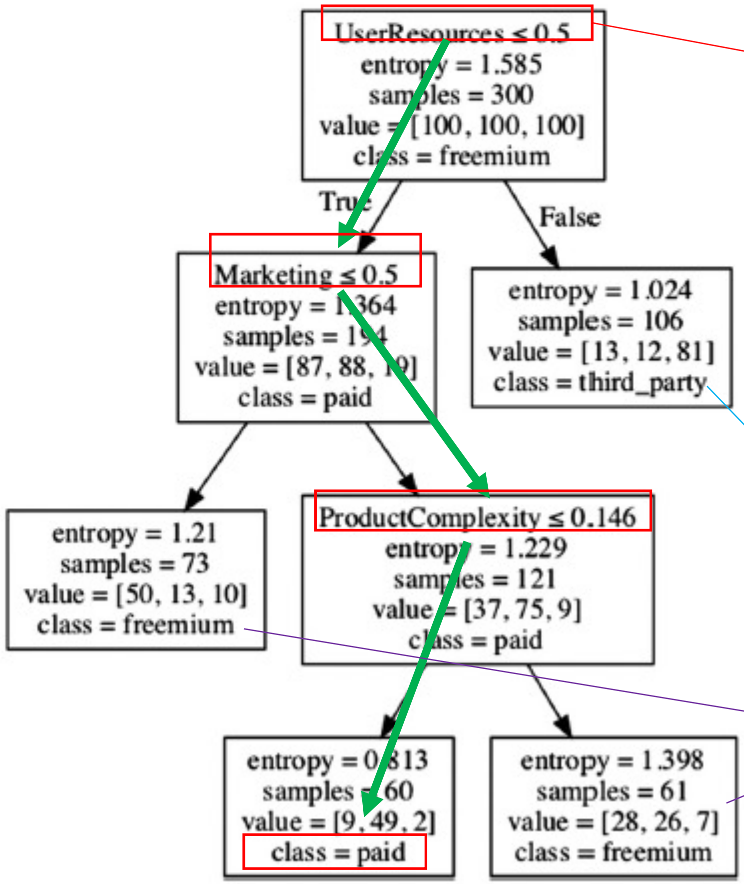


FIGURE 1 Decision tree (popular products): Classifies revenue models. Test accuracy (i.e., percent correctly classified) = 85%. Optimal hyperparameters: Minimum samples at leaf = 60, minimum samples at a split = 5, maximum depth = 3. Training accuracy = 69%. Cross-validation accuracy = 69% (standard deviation of 11%)

假设有一个APP特征如下：

user resources	offline brand	marketing	product complexity	design
0.1	1	1	0.01	1

决策树预测流程如下(见左图绿色箭头)：

1、user resources=0.1 <0.5 True 进入左子树

2、marketing = 1 >0.5 False 进入右子树

3、product complexity=0.01<0.146 True 进入左子树

4、得到结果：paid

发现：

1、位置越靠上的特征越重要

2、third party唯一由user resources决定，因此user resources更匹配third party这种收益模式

3、可以看到freemium分成了两个叶子节点，这与探索性数据分析、案例分析中freemium分成bundled freemium、fragmented freemium一致

方法概述—机器学习

Penalized multinomial logistic regression						
Revenue model	**User resources**	**Marketing**	**Design**	**Offline brand**	**Product complexity**	**Product quality**
Third-party (free)	2.04	0.04	−0.08	−0.53	0.06	−0.20
Paid	−1.07	0.95	0.78	0.58	−0.13	0.19
Freemium	−0.97	−0.99	−0.70	−0.06	0.08	0.02

Random forest	
Feature	**Importance (mean impurity decrease)**
User resources	0.55
Marketing	0.18
Product complexity	0.15
Design	0.06
Offline brand	0.03
Product quality	0.03

Notes: *Penalized multinomial logistic regression*: Test accuracy (i.e., percent correctly classified) = 80%. Coefficients (log-odds) fit on popular products (N = 300). Optimal hyperparameter (inverse of regularization strength), C = 1. Training accuracy = 69%. Cross-validation accuracy = 68% (standard deviation of 12%). *Random forest*: Test accuracy (i.e., percent correctly classified) = 80%. Predictor importance (n_{trees} = 100) analysis of popular products (N = 300). Optimal hyperparameters: Minimum samples at leaf = 5, minimum samples at a split = 40, maximum depth = 3. Training accuracy = 71%. Cross-validation accuracy = 68% (standard deviation of 11%).

假设有一个APP特征如下：

user resources	offline brand	marketing	product complexity	design
0.1	1	1	0.01	1

逻辑回归算法类似多元线性回归，表格内数字为系数，将数据代入计算，三个收益模式的公式中哪个值最大，则属于哪个收益模式

某特征对应系数绝对值越大，该特征影响力越大正号代表正向影响，负号代表负向影响。由此得出：third-party中user resources最重要；Paid中marketing最重要；freemium中product complexity最重要

随机森林算法与前两个算法类似，但文章并没有用到其“预测”这一功能
该算法可以直接给出特征的重要性，这里可以看出对一个popular产品而言，user resources、marketing比较重要

方法概述—机器学习

结论与发现：

1. Freemium可划分为bundled freemium、fragmented freemium
2. 用户资源(user resources)有价值时，适合third-party模式

   产品复杂性(product complexity)越高，越适合freemium模式

   线下品牌(offline brand)、营销(marketing)、设计(design)更适合paid模式

结论与案例研究、探索性数据分析结果是一致的

得出总体结论：**收益模式**需要与**产品商业活动特征**相互匹配

方法概述—机器学习

TABLE 1 Comparison: Multi-case theory building and machine learning

	Multi-case theory building	Machine learning
Definition	Process for finding patterns in data using two or more cases	Algorithms for finding patterns in quantitative data
Objectives	Robust, accurate, and generalizable theory	Robust, accurate, and generalizable prediction
Guards against overfitting	Replication logic	Cross-validation
Guards against excess complexity	Construct abstraction	Regularization techniques
A priori assumptions	Few	Few
Sampling	Theoretical	Balanced random
Theoretical constructs	Identified from data	Selected from among many possible features in data
Final model selection	Researcher	Algorithm
Scale	Very small	Small to very large
Strengths	Empirically grounded Rich description Identification of major theoretical constructs, relationships, and mechanisms	Empirically grounded Large scale Selection of major constructs and identification of precise patterns (e.g., effect size and direction, nonlinearities, equifinality, configurations)
Weaknesses	Small scale Researcher dependent Imprecise identification of patterns (e.g., interactions, nonlinearities, effect sizes)	Cross-sectional Atheoretical, often "black box" No causal inference or significance tests as of yet