2022 年“泰迪杯”数据分析技能赛

B 题

银行客户忠诚度分析

一、背景

目前银行产品存在同质化现象，客户选择产品和服务的途径越来越多，对产品的忠诚度越来越低。为了提高客户对银行的忠诚度和银行营销量，商业银行迫切需要转变经营理念，从“产品销售导向”业务模式向“以客户为中心”转变，为客户带来极致体验和价值成长，形成路径依赖，进而实现价值共赢。

客户忠诚度主要体现为客户的行为和态度。客户行为主要表现为产品重复购买的频率，而客户态度主要表现为情感的倾向。为了有效挖掘客户忠诚度，需要从短期客户产品购买数据和长期客户资源信息中分析客户需求指标。其中，短期客户忠诚度分析是通过产品的购买数据，分析不同指标客户对银行产品的购买依赖度从而提供更好的销售服务；长期客户忠诚度分析则是从客户资源信息数据中挖掘客户流失因素、预测可能流失的客户，尽可能留住高价值客户。

二、数据说明

银行客户忠诚度分析数据包括短期客户产品购买数据和长期客户资源信息数据，附件数据详细情况如表 1 所示。

表 1　附件数据详细情况

数据集	数据名称	备注
短期客户产品购买数据	short-customer-data.csv	客户购买产品的记录数据
长期客户资源信息数据	long-customer-train.csv	训练集数据
	long-customer-test.csv	测试集数据

短期客户产品购买数据记录了往期银行营销活动中客户购买产品的信息，包含客户的基本信息、上次活动后拜访客户信息和上次活动产品购买结果等，具体的数据指标说明如表 2 所示。

表 2　短期客户产品购买数据指标说明

	字段	说明
基本数据	user_id	客户 id，例 BA2200001
	age	年龄（数字）
	job	工作类型包含 11 种，分别为行政人员（admin.），蓝领（blue-collar）、企业家（entrepreneur）、家政（housemaid）、企业管理层（management）、退休（retired）、个体经营者（self-employed）、服务行业人员（services）、学生（student）、技术员（technician）、失业（unemployed）
	marital	婚姻状况包含 3 种，分别为离婚（divorced）、已婚（married）、单身（single），注：离婚指离婚或丧偶）
	education	教育情况包含 5 种，分别为研究生以上（postgraduate）、高中（high school）、文盲（illiterate）、专科（junior college）、大学学位（undergraduate）
	default	信用违约情况包含 2 种，分别为否（no）、是（yes）
	housing	住房贷款情况包含 2 种，分别为否（no）、是（yes）
	loan	个人贷款情况包含 2 种，分别为否（no）、是（yes）
上次活动后拜访客户信息	contact	联系人通信类型包含 2 种，分别为蜂窝（cellular）、电话（telephone）
	month	最近一次拜访客户的月份，分别为一月（jan）、二月（feb）、三月（mar）……十一月（nov）、十二月（dec）
	day_of_week	最近一次拜访客户的星期，分别为星期一（mon）、星期二（tue）、星期三（wed）、星期四（thu）、星期五（fri）
	duration	最近一次拜访客户的通话时长，以秒为单位（数字），如果通话时长=0，表示没有成功联系上客户
其他属性	poutcome	上一次银行活动，客户购买产品的结果包括 3 种，分别为失败（failure）、不存在（nonexistent）、成功（success）
产品购买结果	y	本次银行活动客户购买产品的结果，分别为否（no）、是（yes）

长期客户资源信息数据记录了往期该银行客户流动状态及客户信息，包括客户基本信息、客户户龄与金融资产、客户活跃状态与流失情况等，具体的数据指标说明如表 3 所示。

表 3　长期客户资源信息数据指标说明

序号	字段	说明
1	CustomerId	客户 ID
2	CreditScore	表示信用资格，数值越大表明信用越高
3	Gender	客户性别，0 表示男性，1 表示女性
4	Age	客户年龄
5	Tenure	账号户龄，客户在这家银行存款的时长，以年为单位
6	Balance	AUM，客户的金融资产
7	NumOfProducts	客户购买产品数量
8	HasCrCard	客户持有信用卡状态，客户有信用卡为 1，否则为 0
9	IsActiveMember	客户活动状态，客户处于活跃状态为 1，否则为 0
10	EstimatedSalary	客户个人年收入
11	Exited	客户流失情况，已流失为 1，否则为 0

三、目标

（1） 对客户数据进行预处理，并对字符型数据进行特征编码。

（2） 基于短期客户产品购买数据，分析不同指标客户对银行产品的购买依赖度，并进行可视化呈现。

（3） 基于长期客户资源信息数据，分析客户流失因素，并进行可视化呈现。

（4） 依据长期客户资源信息数据的分析结果构建相关指标，对银行客户长期忠诚度进行预测。

四、任务

任务 1　数据探索与清洗

分别对短期客户产品购买数据“short-customer-data.csv”（简称短期数据）和长期客户资源信息数据的训练集“long-customer-train.csv”（简称长期数据）进行数据探索与清洗。

任务 1.1　数据探索与预处理

(1) 探索短期数据各指标数据的缺失值和“user_id”列重复值，删除缺失值、重复值所在行数据。请在报告中给出处理过程及必要结果，完整的结果保存到文件“result1_1.xlsx”中。

(2) 长期数据中的客户年龄“Age”列存在数值为-1、0 和“-”的异常值，删除存在该情况的行数据；“Age”列存在空格和“岁”等异常字符，删除这些异常字符但须保留年龄数值，将处理后的数值存于“Age”列。请在报告中给出处理过程及必要结果，完整的结果保存到文件“result1_2.xlsx”中。

任务 1.2　对短期数据中的字符型数据进行特征编码，如将信用违约情况{‘否’,‘是’}编码为{0,1}。请在报告中给出处理思路、过程及必要结果，完整的结果保存到文件“result1_3.xlsx”中。

任务 2　产品营销数据可视化分析

基于短期数据分析不同指标客户与购买银行产品行为的关联性，挖掘短期客户对银行的忠诚度。

任务 2.1　计算短期数据所有指标之间的相关性，绘制相关系数热力图，并在报告中对结果进行必要分析。

任务 2.2　在同一画布中，绘制反映两种产品购买结果下不同年龄客户量占比的分组柱状图，*x* 轴为年龄，*y* 轴为占比数值，并在报告中对结果进行必要分析。

任务 2.3　在同一画布中，绘制蓝领（blue-collar）与学生（student）的产品购买情况饼图，并设定饼图的标签，显示产品购买情况的占比。

任务 2.4　以产品购买结果为 *x* 轴、拜访客户的通话时长为 *y* 轴，绘制拜访客户的通话时长箱线图，并在报告中对结果进行必要分析。

任务 3　客户流失因素可视化分析

基于长期数据分析导致银行客户流失的因素，并进行可视化呈现。

任务 3.1　在同一画布中，绘制反映两种流失情况下不同年龄客户量占比的折线图，*x* 轴为年龄，*y* 轴为占比数值。

任务 3.2　在同一画布中，绘制反映两种流失情况下客户信用资格与年龄分布的散点图，*x* 轴为年龄，*y* 轴为信用资格。

任务 3.3　构造包含各账号户龄在不同流失情况下的客户量占比透视表（详见表 4），并在同一画布中绘制反映两种流失情况的客户各账号户龄占比量的堆叠柱状图，*x* 轴为客户的户龄，*y* 轴为占比量。

表 4　透视表

Tenure / Exited	0	1	2	3	4	5	6	7	8	9	10
0											
1											

注：Tenure 中的 0，…，10 表示客户的户龄，而 Exited 中的 0 表示客户未流失，1 表示已流失。

任务 3.4　新老客户各资产阶段的客户流失情况分析。

(1) 按照表 5 和表 6 对账号户龄和客户金融资产进行划分，并分别进行特征编码作为新的客户特征，其中客户状态存于“Status”列，资产阶段存于“AssetStage”列，编码结果保存到文件“result3.xlsx”中。

表 5　账号户龄划分情况

账号户龄区间	客户状态
[0, 3]	新客户
(3, 6]	稳定客户
>6	老客户

表 6　客户金融资产划分情况

客户金融资产区间	资产阶段
[0, 50000]	低资产
(50000, 90000]	中下资产
(90000, 120000]	中上资产
>120000	高资产

(2) 统计新、老客户在各资产阶段中流失的客户量，在同一画布中绘制热力图，热力图颜色的最大和最小取值设为 1300 和 100，并在报告中对结果进行必要分析。

任务 4　特征构建

基于长期数据提取影响客户流失的因素，构建与银行客户长期忠诚度相关的特征，将结果保存到文件“result4.xlsx”中。

(1) 根据表 7，构建新老客户活跃程度的特征，并将结果存于“IsActiveStatus”列。

表 7　新老客户活跃程度特征构建规则

新老客户活跃程度		活跃状态	
		0	1
账号户龄	新客户	0	3
	稳定客户	1	4
	老客户	2	5

(2) 根据表 8，构建不同金融资产客户活跃程度的特征，并将结果存于“IsActiveAssetStage”列。

表 8　不同存款额客户活跃程度特征构建规则

不同金融资产客户活跃程度		活跃状态	
		0	1
资产阶段	低资产	0	6
	中下资产	1	7
	中上资产	2	8
	高资产	3	9

(3) 根据表 9，构建不同金融资产信用卡持有状态的特征，并将结果存于“CrCardAssetStage”列。

表 9　不同金融资产信用卡持有状态特征构建规则

不同金融资产信用卡持有状态		信用卡持有状态	
		0	1
资产阶段	低资产	0	6
	中下资产	2	7
	中上资产	5	9
	高资产	5	9

任务 5　银行客户长期忠诚度预测建模

长期数据存在“Exited”特征分布不均衡、各项数值分布跨度大等现象。体现为：未流失客户量是已流失客户量的 3 倍以上；客户信用资格最大数值达到 25 万，而客户活动状态则为 0 和 1 等。考虑上述现象，对银行客户长期忠诚度进行预测。

(1) 选取适当的客户特征，建立客户长期忠诚度预测模型。客户特征可以从客户信用资格、性别、年龄、户龄、金融资产、客户购买产品数量、持有信用卡状态、活动状态和个人年收入等指标中直接选取，也可以参照任务 4 构建。在报

告中给出特征选取的依据、建立预测模型的思路和过程。

(2) 使用混淆矩阵、F1 Score 等方法对预测模型进行评估，在报告中给出评估的方法和结果。

(3) 对“long-customer-test.csv”测试数据进行预测，将全部预测结果以表 10 形式保存为文件“result5.xlsx”，其中 0 表示客户没有流失，1 表示客户流失。并将表 11 中的 5 个客户 ID 的预测结果在报告中列出。

表 10　result5.xlsx 预测结果

CustomerId	Exited
15000001	0
15000002	1
……	……

表 11　指定的 5 个客户 ID 的预测结果

CustomerId	Exited
15579131	
15674442	
15719508	
15730076	
15792228	

五、关于竞赛成果提交的说明

1. 登录方式

请使用**队长**的账号登录数睿思网站（www.tipdm.org），进入第五届技能赛页面。为保证成功提交，**请使用谷歌浏览器无痕模式**。

2. 作品提交

报告以 PDF 格式提交，文件名为“**report.pdf**”，要求逻辑清晰、条理分明，内容包括每个任务的完成思路、操作步骤、必要的中间过程、任务的结果及分析。

3. 附件提交

3.1　将任务 1、任务 2、任务 3、任务 4、任务 5 的**源程序**存放到“**program**”文件夹中。

3.2　将结果文件“**result1_1.xlsx**”、“**result1_2.xlsx**”、“**result1_3.xlsx**”、

“**result3.xlsx**”、“**result4.xlsx**”、“**result5.xlsx**”存放到“**result**”文件夹中。

3.3 将任务 2、任务 3 的可视化图片保存到“**image**”文件夹中。

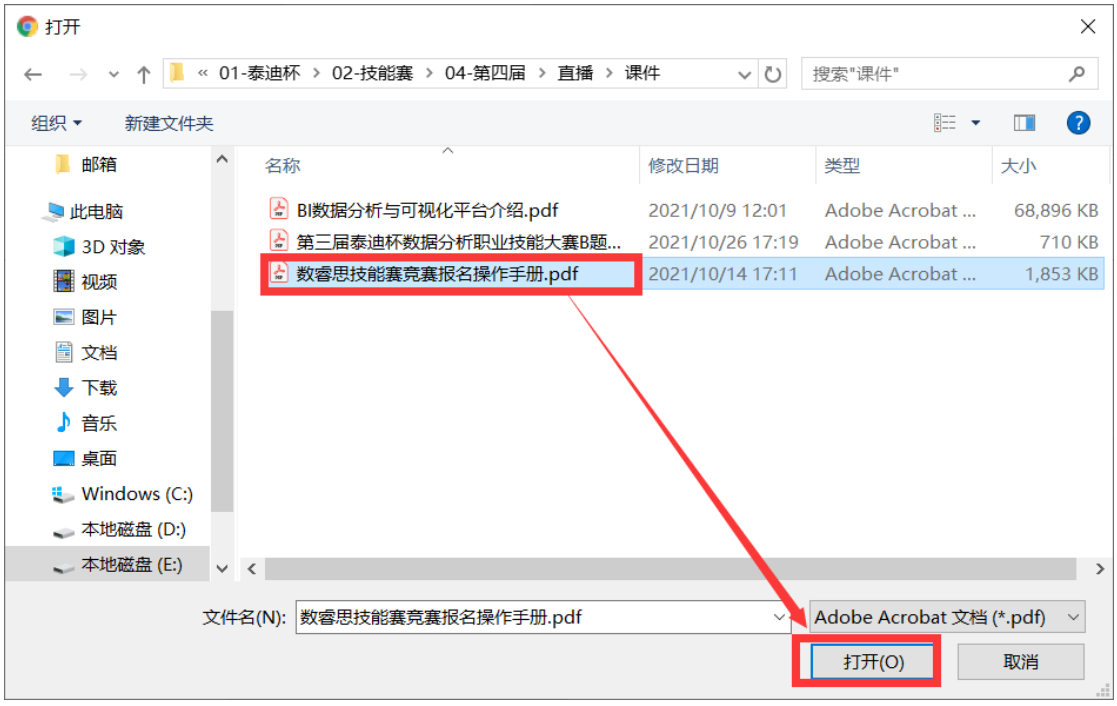
3.4 将程序文件夹“**program**”、结果文件夹“**result**”、可视化图片文件夹“**image**”以及报告的 **Word 版本**打包成“**appendix.zip**”，作为附件提交。

4. 提交界面

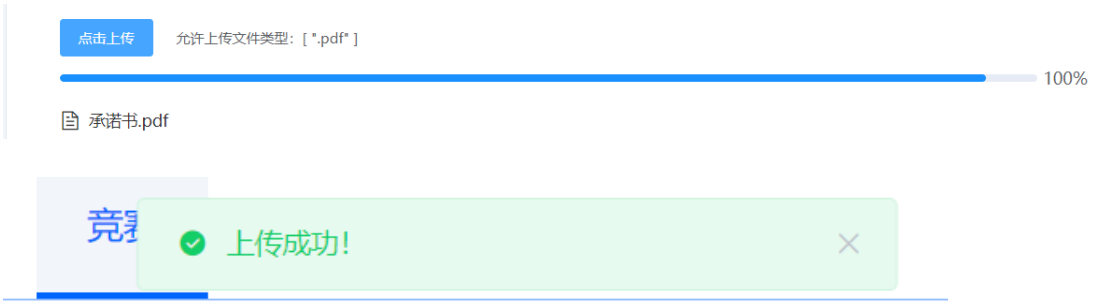
4.1 找到赛题提交入口。

4.2　点击“点击上传”按钮。

4.3　选择需要上传的对应文件，点击“打开”。



4.4　进度条加载完成后会有“上传成功”提示。



4.5　页面如下图即为上传提交成功，多次提交会以最后一次为准。