金融服务机构的资金流量预测

1 项目背景

某金融服务机构拥有上亿会员,并且业务场景中每天都涉及大量的资金流入和流出,面 对如此庞大的用户群,资金管理压力会非常大。在既保证资金流动性风险最小,又满足日常 业务运转的情况下,精准地预测资金的流入流出情况变得尤为重要。企业希望能精准预测未 来每日的资金流入流出情况。对货币基金而言,资金流入意味着申购行为。资金流出意味着 赎回行为。

2 项目目标

基于企业希望精确预测资金流入流出数量的需求,设定项目目标为:预测蚂蚁金服次月 每天的申购总额。

3 项目步骤

3.1 工程前期准备

3.1.1 导入数据

(1) 介绍数据

用户申购赎回数据表:表中包含 2013 年 7 月 1 日 至 2014 年 8 月 31 日的申购和赎回 信息、以及所有的子类目信息。数据经过脱敏处理,脱敏之后的数据,基本保持了原数据分 布。数据主要包括用户操作时间和操作记录,其中操作记录包括申购和赎回两个部分。金额 的单位是分,即 0.01 元人民币。 如果用户今日消费总量为 0,即 consume_amt=0,则四个 子类目为空。如表 3-1 所示。

表 3-1 用户申购赎回数据表

属性	含义	示例
user_id	用户 id	1234
report_date	日期	20140407
tBalance	今日余额	109004
yBalance	昨日余额	97389
total_purchase_amt	今日总购买量 = 直接购 买 + 收益	21876
direct_purchase_amt	今日直接购买量	21863
purchase_bal_amt	今日支付宝余额购买量	0
purchase_bank_amt	今日银行卡购买量	21863
total_redeem_amt	今日总赎回量 = 消费 + 转出	10261
consume_amt	今日消费总量	0
transfer_amt	今日转出总量	10261
tftobal_amt	今日转出到支付宝余额总量	0
tftocard_amt	今日转出到银行卡总量	10261
share_amt	今日收益	13
category1	今日类目 1 消费总额	0
category2	今日类目 2 消费总额	0
category3	今日类目 3 消费总额	0
category4	今日类目 4 消费总额	0

(2) 上传数据到 Python 数据挖掘建模平台

在新增数据源上,选择本地上传数据,如图1所示。

	Py	/thon数据	挖掘建模	平台					Ţ	•	0	
▲ 首页		我的数据源	共享数据源	<u>9</u>								
数据源	[十 新增数	据源 ▲			请输入表名	选择状态 🔻 请说	超早创建时间		C	搜索	2
「工程	[■ 数据来源	于文件	创建人	数据来源	同步状态	创建时间	操作				
● 个人组件		● 数据来源	于数据库									
₩ 模型												
日子						暂无数据						

图 1 本地上传数据源

在本地路径上选择文件,填写在平台新建的目标表名,如图 2 所示。

	新闻的数据研究	×
1 文件属性	2 预览数据	3 字段设置
• 新建目标表名	被称文件 total_purchase_amt	
列分隔符	逗号 (.) 🔹 文件编码 UTF-8 👻	
存储有效期 (天)	180 一 十 预览设置 分页显示 🔻	
		重置下一步

图 2 本地选择文件上传

根据文件的数据,可以修改文件的字段名和类型,如图 3 所示。

				新	建数据	訪原					×
1					2)					3
文件属性				1	预览数	据					字段设置
注意: 【字段名】只能是 原字段	以字母开头,由小写英文字母 字段名	、 数字、下划线组成 类型		ŧ	度		程	镀		备注	
names	names	数值	-	255	-	+	0		+		
report_date	report_date	日期	-	255	-	+	0		+		
total_purchase_amt	total_purchase_amt	数值	-	255	-	+	0		+		
											上一步 确定

图 3 字段设置

上传成功,可以在平台的数据源上查看数据,单击数据源操作的查看按钮如图 4 所示, 数据预览如图 5 所示。

+ 新增数据源 ▼				请输入表名	选择	¥状态 ▼	请选择创建时间	
表名	创建人	数据来源	同步状态		创建时间		操作	
total_purchase_amt	xinyou	结构化文件	同步完成		2019-05-28 09:59:52		 	•
credict_card	xinyou	结构化文件	同步完成		2019-05-28 09:29:59		۰ 🛍 🔹	¢
discdata	xinyou	结构化文件	同步完成		2019-05-28 08:46:49		۰ ش	¢
hotspotdata	xinyou	结构化文件	同步完成		2019-05-27 15:33:36		۵ <u>۱</u> •	•
user_dat	xinyou	结构化文件	同步完成		2019-05-27 13:59:09		@ 🛍 •	•

图 4 单击预览数据按钮

			预览数据 (分页加载)
names	report_date	total_purchase_a mt	
1	2013-07-01	32488348	
2	2013-07-02	29037390	
3	2013-07-03	27270770	
4	2013-07-04	18321185	
5	2013-07-05	11648749	
6	2013-07-06	36751272	
7	2013-07-07	8962232	
8	2013-07-08	57258266	
		共	427条 100条页 💌 🤇 1 2 3 4 5 〉 前往 1 页

图 5 数据预览

3.1.2 新建空白工程

右击我的工程,新建一个空白的工程,如图 6 所示。

	Python数据挖掘建模	评台	4 8 9 0
▲	工程 〇	🛓 🔍 🔍 120%) % 🖺 🗅 O	工程信息
	 ▶ #044-740 ● 新建工程 ▲ 导入工程 ▲ 导出工程 ● 汚し工程 		未选择工程
⑦ 个人组件	■ 添加文件夹 组件		
▶ 模型 旧日 任务	输入内容进行过滤		
	模型		

图 6 新建工程

填写工程的信息,包括工程名称和工程描述,如图 7 所示。

	创建工程	×
* 工程名称	金融服务机构的资金流量预测	
工程描述	某金融服务机构拥有上亿会员,并且业务场景中每天都涉及大量的资金流入和流出,面对如此庞大的用户群,资金管理压力会非常大。在既保证资金流动性风险最小,又满足日常业务运转的情况	×
工程位置	▼我的工程	
	重置 确定	

图 7 填写工程信息

3.2 数据预处理

读取 total_purchase_amt 数据,步骤如图 8 所示。

- (1)选择工程。
- (2) 选择输入源组件。
- (3) 拖入输入源组件。
- (4) 填写数据表名。
- (5) 单击更新按钮,更新出数据。

	📩 Q Q 120% % 🖺 O	✓ 字段属性
▲ 竞赛网站用户		数据表
▲ 气象与输电线	11111111111111111111111111111111111111	total_purchase_amt
▲ 应用系统负载		
▲ 信用卡高风险		字段信息
▲ 金融服务机构		
组件		字段 类型 取值范围
输入内容进行过滤		names 数值
 ▼ 新久/輸出 		report_dat e 日期
≓ 输入源		total_purc base amt 数值
▶ 预处理	•	
▶ 统计分析		
▶分类		
▶ 聚类		
▶ 时序模型		
▶ 关联规则		
################################		
横型		> 组件描述

图 8 输入源组件

3.2.1 缺失值处理

数据可能存在缺失值,先对数据进行缺失值处理,步骤如错误!未找到引用源。所示。 (1) 找到预处理→缺失值处理组件。

(2) 拖入缺失值处理组件,并将输入源和缺失值处理组件连接。

(3) 单击更新按钮,勾选全部数据的字段作为输出字段。

(4) 对缺失值处理组件右键,选择运行该节点。

or type of advances to the	🗻 Q Q 120% % 🖹 O	~ 字詞	2属性		
▲ 竞赛网站用户		特征			
▲ 气象与输电线	11111111111111111111111111111111111111	6			
▲ 应用系统负载		医加热	的计读字体由		
▲ 信用卡高风险	714-/25-LLTH	Contrast 1	PAGE PD 7 1911		
▲ 金融服务机构			字段	类型	取值范围
8/4			names	数值	
财排库			report_date	日期	
■数据筛选			total_purchas	数值	
翼 分组聚合			e_am		
翼 修改列名					
X Python脚本					
X 缺失值处理					
■ 数学类函数					
其 特征构造					
■ 数据离散化					
国 标准化数据还原					
🗙 数据标准化					
■数据编码化		> 参数	女设置		
		N	14.489.140		

图 9 缺失值处理组件

(5)运行完成后,对缺失值处理组件右键,选择查看数据,如错误!未找到引用源。所示。

	预览数据	
names	report_date	total_purchase_amt
1	2013-07-01	32488348
2	2013-07-02	29037390
3	2013-07-03	27270770
4	2013-07-04	18321185
5	2013-07-05	11648749
6	2013-07-06	36751272
7	2013-07-07	8962232
8	2013-07-08	57258266
共 427 条 25 条/页 ▼	<pre>1 2 3 4 5 6</pre>	••• 18 > 前往 1 页

图 10 缺失值处理结果

3.2.2 平稳性检验

选择平稳性检验,步骤如图 11 所示。

(1) 找到统计分析→平稳性检验。

(2) 拖入平稳性检验组件,将缺失值处理和平稳性检验组件连接。

(3) 单击更新按钮,时序特征勾选 total_purchase_amt 字段作为检验字段。

(4) 对平稳性检验组件右键,选择运行该节点。

工程 〇	🛓 🔍 🔍 120% % 🖺 🗅 🖸	◇ 字段属性			
▲ 竞赛网站用户	(a) 100	数据			0
▲ 气象与输电线					
▲ 应用系统负载		字段	类型	取值范围	
▲ 16用卡筒风腔 A 全融服务机构		namee	(約)店		
- IIIR0000757011-5	設定になっていた。	report dat	9,11H		
组件		e	日期		
Mann-Whitney.		total_purc	教信		
6 K-S检验		hase_amt	~		
● 主表机1	25 25 25 25 25 25 25 25 25 25 25 25 25 2				
© 方差齐性检验		时序特征			0
& Granger因果					Ŭ
• 主成分分析		total purchase	omt		
6 频数统计		total_purchase	_am		
✿ 因子分析					
6 半稳性检验					
▶ 万 <u>只</u> ▶ 回归					
▶ 聚类					
模型		> 组件描述			

图 11 平稳性检验组件

(5)运行完成后,对平稳性检验组件右键,选择查看报告,平稳性检验的报告结果如图 12

所示。

算法运行报告	×
平稳性检验结果	
检验结果	
Test statistic: -1.5898802926313504 p-value: 0.4886749751375929 Number of lags used: 18 Number of observations used for the ADF regression and calculation of the critical values: 408 Critical values for the test statistic at the 5 %: -2.8686500930967354 Critical values for the test statistic at the 1 %: -3.446479704262724 Critical values for the test statistic at the 1 %: -3.47677574605	l
ACF	
• • • • • • • • • • • • • • • • • • •	
	下载

图 12 平稳性检验报告

3.2.3 纯随机性检验

选择平稳性检验,步骤如图 13 所示。

- (1) 找到统计分析→纯随机性检验。
- (2) 拖入平稳性检验组件,将缺失值处理和纯随机性检验组件连接。
- (3) 单击更新按钮,特征勾选 total_purchase_amt 字段作为检验字段。
- (4) 对纯随机性检验组件右键,选择运行该节点。

	📩 🔍 🔍 120% % 🖺 🖸	∨ 字段属性			
▲ 竞赛网站用户		数据			
▲ 气象与输电线	11111111111111111111111111111111111111	0			
▲ 应用系统负载		字段	迷刑	取債范围	
▲信用卡高风险			~		
▲ 金融服务机构	致 缺失值处理	names	数值	-	
组件		report_dat e	日期		
C Mann-Whitney C K-S检验		total_purc hase_amt	数值		
€ 全表统计	● 平稳性检验				
€ 纯随机性检验					
方差齐性检验		特征			0
Granger因果…	40000000	0			
主成分分析	A DURY PLAN ALL	total nurchase	amt		
€ 频数统计		total_purchase	_ann		
B子分析					
《 平稳性检验					
▶ 分类					
) oli					
▶ 聚类					
模型		> 组件描述			

图 13 纯随机性检验组件

(5)运行完成后,对纯随机性检验组件右键,选择查看报告,纯随机性检验的报告结果 如图 14 所示。

		如果p值小于0.05时,可以证明通过白噪声检验!
lags	pvalue	
1	2.0420594779177276e-65	
2	6.720778898589508e-106	
3	8.148894825818178e-137	
4	1.517669826094083e-165	
5	1.355187994202336e-195	
6	1.372153619425153e-230	
7	6.170102293269174e-269	
-		

图 14 纯随机性检验报告

3.3 模型构建

3.3.1 ARIMA 算法

选择 ARIMA 算法模型,步骤如图 15、图 16 所示。

(1) 找到时序模型→ARIMA 组件。

(2) 拖入 ARIMA 组件,将生成训练数据和 ARIMA 组件连接。

(3) 选择字段属性,单击更新数据,时序列勾选 total_purchase_amt 字段,时间列勾选 report date 字段。

(4) 选择参数设置,设置预测周期数的值为 10,设置自回归项数 p 的值为 7,设置差分 次数 d 的值为 1,设置移动平均项数 q 的值为 2。



图 15 ARIMA 组件_字段属性



图 16 ARIMA 组件_参数设置

(5) 运行完成后,对 ARIMA 组件右键,选择查看报告, ARIMA 的报告如图 17 所示。

		算法运行报告
		模型參数
		需要配置的参数及其取值如下。
参数名称	参数值	
自回归项数p	7	
差分次数d	1	
移动平均项数q	2	
		模型具体信息
		模型具体信息如下。 Statespace Model Results

图 17 ARIMA 的报告