天池竞赛：资金流入流出预测

Author: Haorui HE
Github: HarryHe11/TIANCHI-Purchase-Redemption-Forecast-Challenge
2022.1.29

一、 背景与简介

货币基金是聚集社会闲散资金，由基金管理人运作，基金托管人保管资金的一种开放式基金，专门投向风险小的货币市场工具，区别于其他类型的开放式基金，具有高安全性、高流动性、稳定收益性，具有“准储蓄”的特征。

在大数据时代背景下，蚂蚁金服等拥有上亿用户的互联网金融公司的真实业务场景中，每天都涉及巨额的资本流入（即货币基金申购）和流出（即货币基金赎回）。面对海量用户群带来的不确定资金流量，互联网金融公司面临着巨大的资金管理压力。如何在满足日常业务运转的同时，最大程度的减小资金流动性风险，成为了此类公司关注的重难点问题。

因此，精准地预测资金的流入流出情况变得尤为重要。在本项工作中，我期望能够通过对余额宝用户的历史申购赎回数据、用户个人信息数据和银行利率波动等业务相关数据的把握，实现对未来 30 天内每日的资金流入流出情况的预测。

二、 问题阐述

我将资金的流入流出预测视作机器学习中的回归问题。令$\boldsymbol{x_i}$表示某日的资金流动相关数据特征向量，其由 n个数据特征$\{f_0, f_1, \ldots, f_{n-1}\}$组成； $\mathbf{y_i} = \{y_{i,0}, y_{i,1}\}$表示当日的资金流入和流出总额，其中$y_{i,0}$表示资本流入总额（即货币基金申购），$y_{i,1}$表示资金流出总额（即货币基金申购）。给定某一段时间内每天的数据特征向量$\boldsymbol{x_i}$，我期望精准地预测该段时间内每天的资本流入总额 $p_{i,0}$和资金流出总额 $p_{i,1}$，使 $p_{i,0}$及 $p_{i,1}$与真实值 $y_{i,0}$及 $y_{i,1}$的相对误差 ε_0和 ε_1尽可能的小，此处的误差计算公式为：

$$\varepsilon_0 = \frac{|y_{i,0} - p_{i,0}|}{y_{i,0}} \tag{1}$$

$$\varepsilon_1 = \frac{|y_{i,1} - p_{i,1}|}{y_{i,1}} \tag{2}$$

申购预测得分 s_0与 ε_0相关，赎回预测得分 s_1与 ε_1相关, 误差与得分之间的计算公式未公布，但该计算公式单调递减，误差越小，得分越高。最终得分 finalScore 的计算公式为：

$$\text{finalScore} = 45\%\, s_0 + 55\%\, s_1 \tag{3}$$

三、 探索性数据分析

（1）时序数据可视化分析

a. 申购和赎回总量时序可视化分析

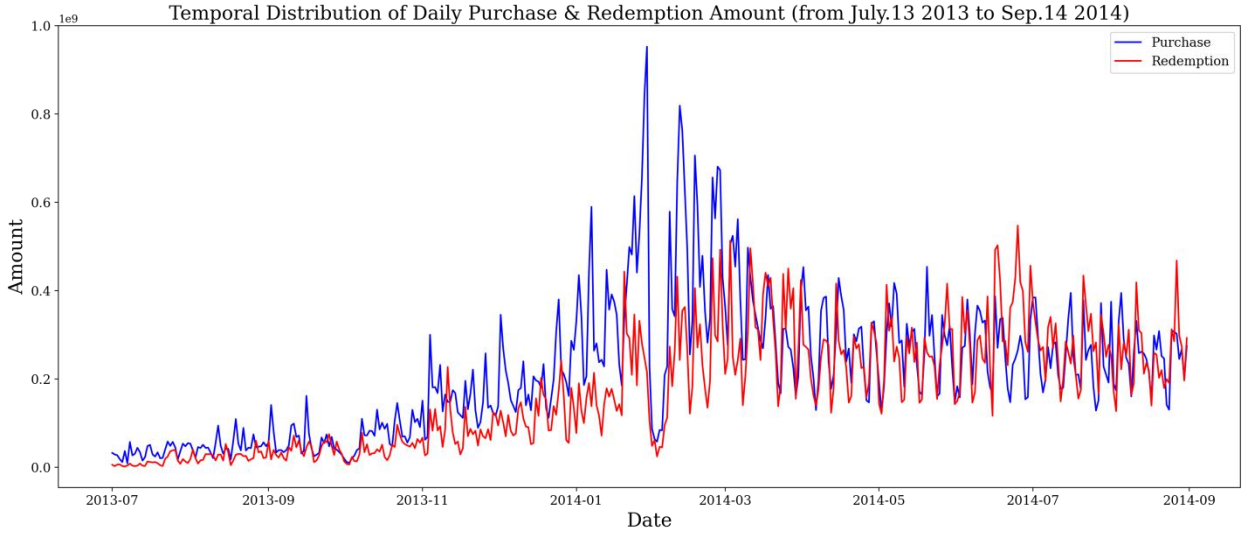

图 1 2013 年 7 月 13 日-2014 年 9 月 14 日每日申购赎回总量

图 1 展示了 2013 年 7 月 13 日到 2014 年 9 月 14 日每日申购和赎回总量。从图中可以看出，2013 年的每日申购和赎回的总量整体低于 2014 年水平。并且，2014 年的第一季度内申购和赎回的每日总量波动幅度较大，可能对模型精度产生影响，特征工程时可以考虑训练集开始日期设置在 2014 年 4 月 1 日。

b. 周内差异可视化分析

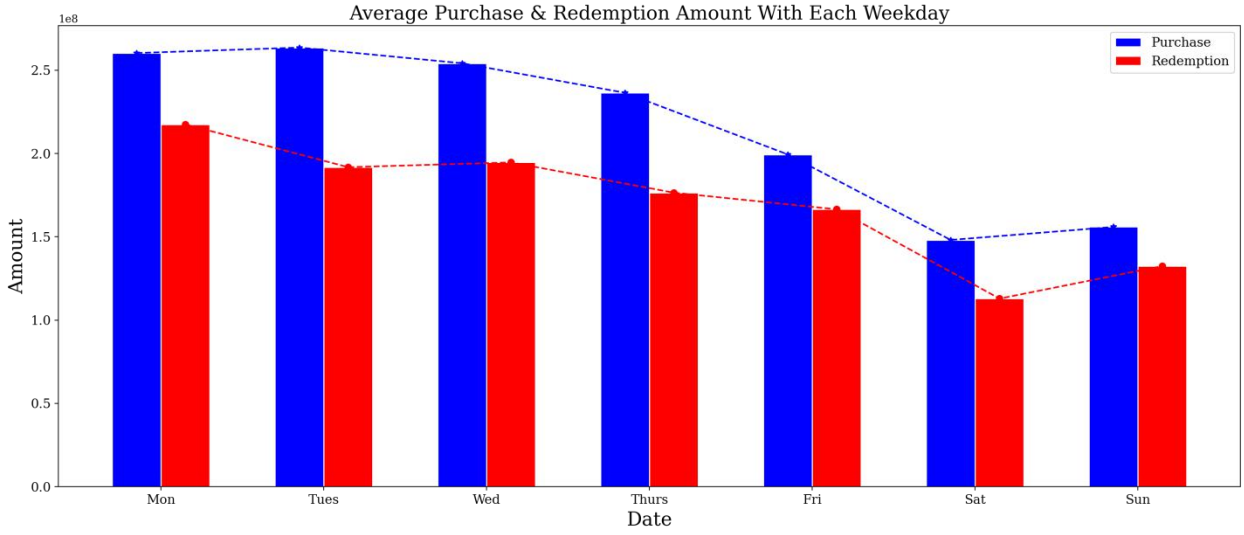

图 2 周内平均申购赎回总量

图 2 展示了周内每日平均申购总量和赎回总量的差异。从图中可以看出，每日平均申购总量和赎回总量呈现了相同的趋势：周一时平均资金流动总量最高，周一到周六资金流动量递减，并于周日少量回暖。除此之外，可以看出周末两天的平均资金流动量低于工作日，因此可以在特征工程时考虑针对当日是否为周末进行特征构建。

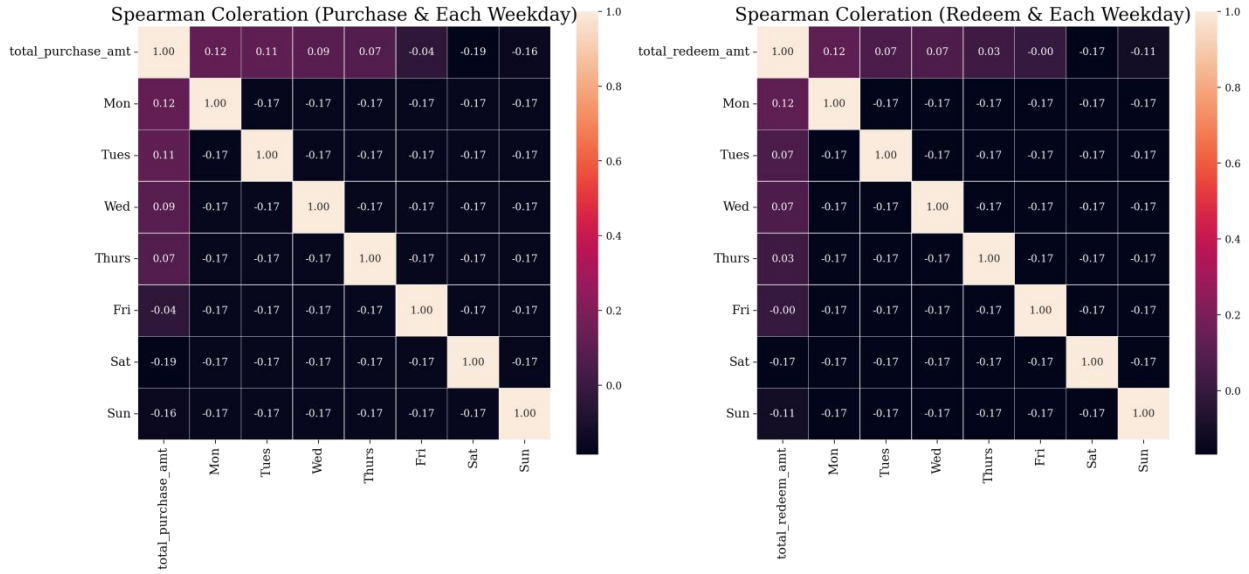

图 3 星期特征与申购赎回总量的 Spearman 相关度

图 3 展示了星期特征与申购赎回总量的 Spearman 相关系数矩阵。由于星期特征是定序变量，申购赎回总量是一个定距变量，因此采用 Spearman 相关系数，但由图中可以看出星期特征与申购赎回总量的线性相关性均较低。

c. 月内差异可视化分析

图 4 展示了月内每日平均申购总量和赎回总量的差异及其频率直方图。从图中可以看出每个月 11 号和 25 号左右申购平均总量较高，25 号左右赎回总量同样较高，而每月初的 1 号和月末 31 号的资金流入流出量都相对较低。由右侧频率直方图可以看出，平均赎回总量的波动相对平均申购量更高。

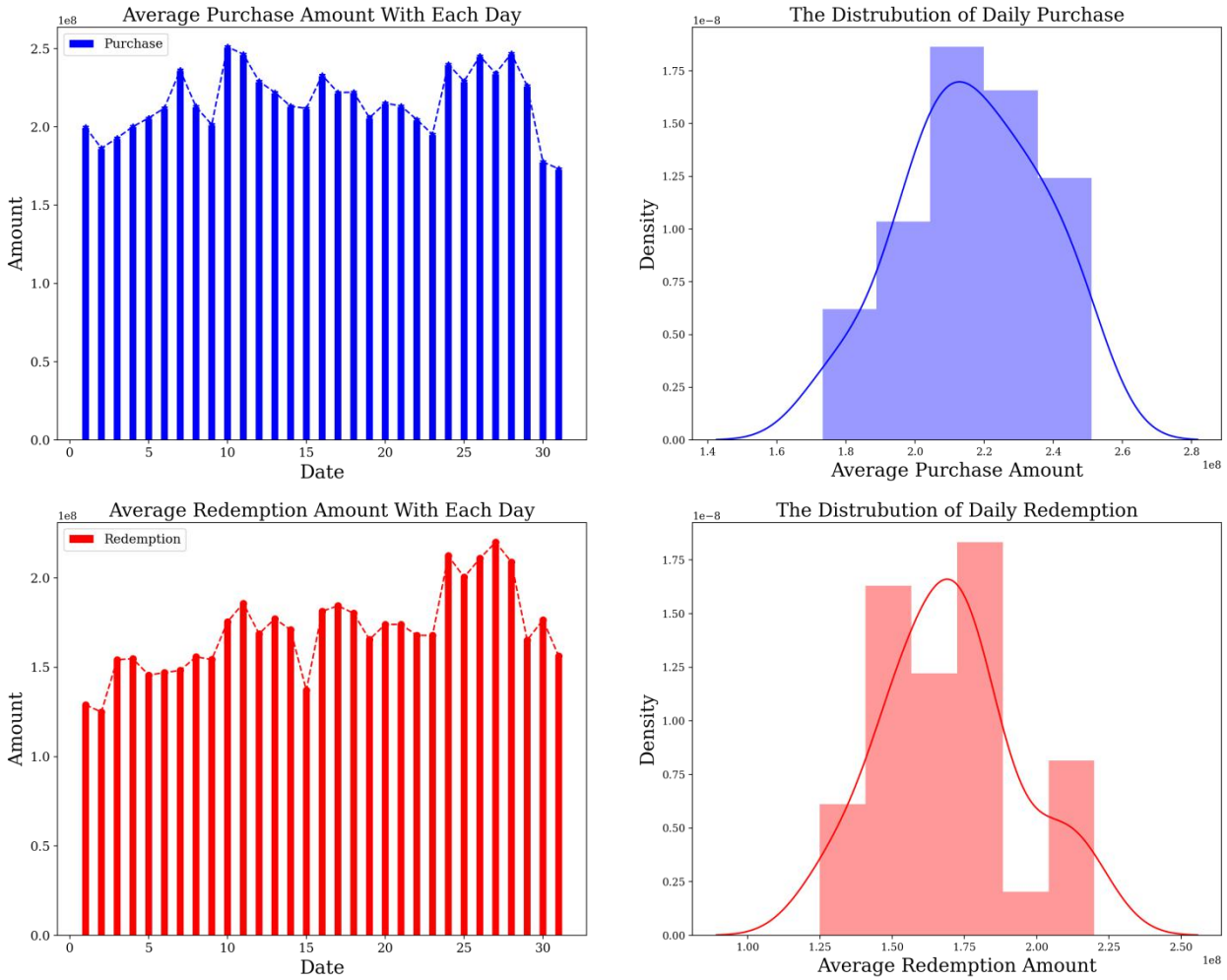

图 4 月内每日平均申购总量和赎回总量的差异及其频率直方图

d. 每月差异可视化与分析

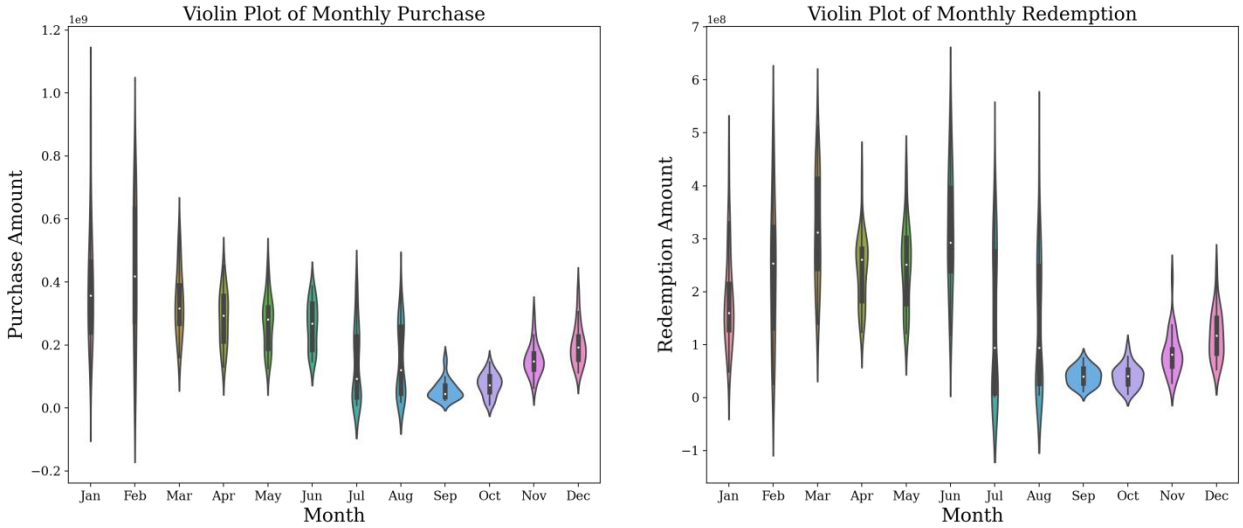

图 5 每月申购总量和赎回总量的小提琴图

图 5 展示了每月的申购总量和赎回总量的小提琴图。从图中可以看出，申购总量和赎回总量都在一月和二月波动较大，分布差异大，而在九到十二月波动较小，分布稳定。赎回总量在三到八月波动比申购总量更大，同时，下半年的资金流动总量中位数低于上半年，可以考虑利用每月平均申购和赎回总量中位数构建特征。

e. 两年同期数据可视化分析（以 2013 年和 2014 年 8 月同期为例）

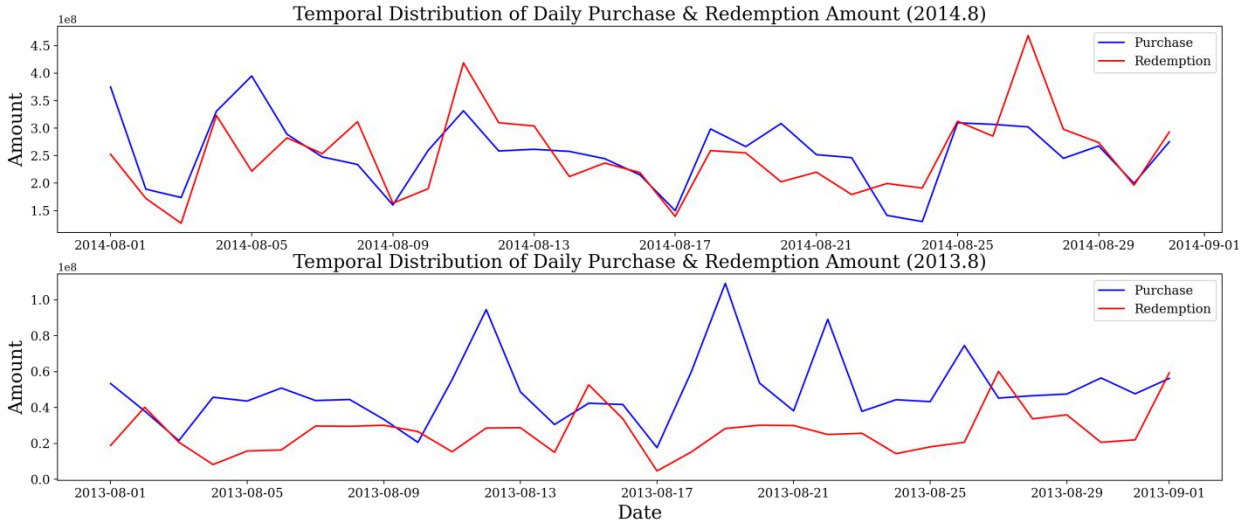

图 6 2013 年 8 月和 2014 年同期每日申购赎回总量

图 6 展示了 2013 年 8 月和 2014 年同期每日申购赎回总量比较。但从图中并不能看出两年同期有明显的趋势相关性。

f. 对节假日与特殊日期（互联网消费节）的可视化分析

国家法定节假日	2013 年	2014 年
元旦	1.1-1.3	1.1-1.3
春节	2.9-2.15	1.31-2.6
清明	4.4-4.6	4.5-4.7
劳动	4.29-5.1	5.1-5.3
端午	6.10-6.12	5.31-6.2
中秋	9.19-9.21	9.6-9.8
国庆	10.1-10.7	10.1-10.7
互联网消费节		
618	6.18	6.18
双十一	11.11	11.11
双十二	12.12	12.12

表 1 2013 年与 2014 年节假日与互联网消费节具体日期

表 1 统计了 2013 年与 2014 年节假日与互联网消费节的具体日期，我期望通过比较这些时间的资金流动总量与平时的差异，观察节假日与互联网消费节对申购和赎回总量的影响。

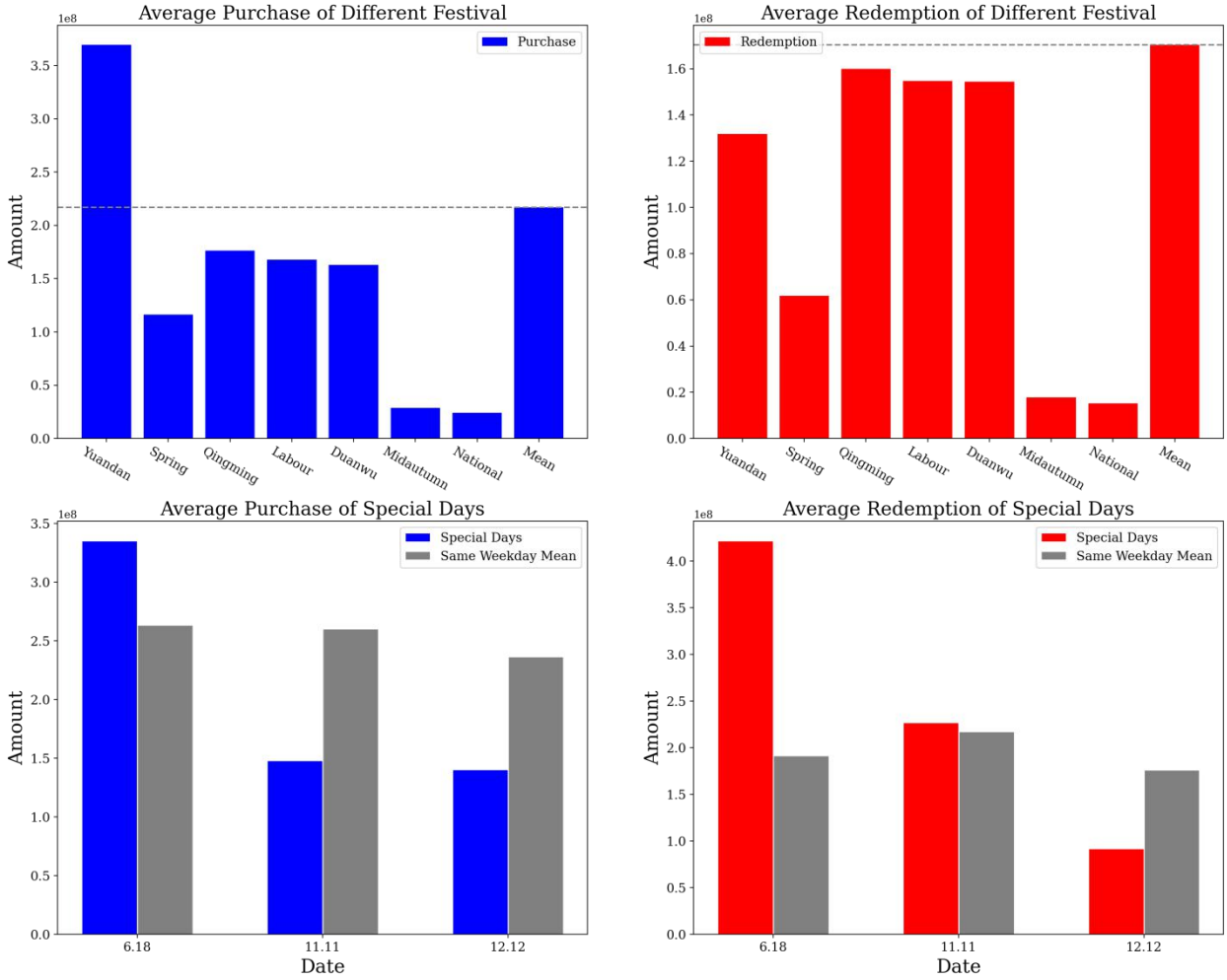

图 7 节假日及互联网消费节与平均申购赎回总量比较

图 7 展示了节假日及互联网消费节与平均申购赎回总量的比较柱状图。从图中可以看出，除了元旦节的申购总量比每日平均申购总量更高之外，其他节日的申购和赎回总量均低于平均，其中中秋节和国庆节的资金流动平均值最低。互联网消费节的资金流动情况也与每周的同天有一定差异，但考虑到测试集时间段中并不包含这三类消费节，该项信息可以考虑忽略。

（2）用户交易和消费行为的可视化分析

a. 大额交易与小额交易行为可视化分析

图 8 展示了周内每日大额交易（交易额大于等于一万元）和小额交易（交易额小于等于一万元）比较箱型图。从图中可以看出，每日大额申购中周内每天都出现了一定异常值。在周末，大小额的申购和赎回总额的中位数均低于工作日。

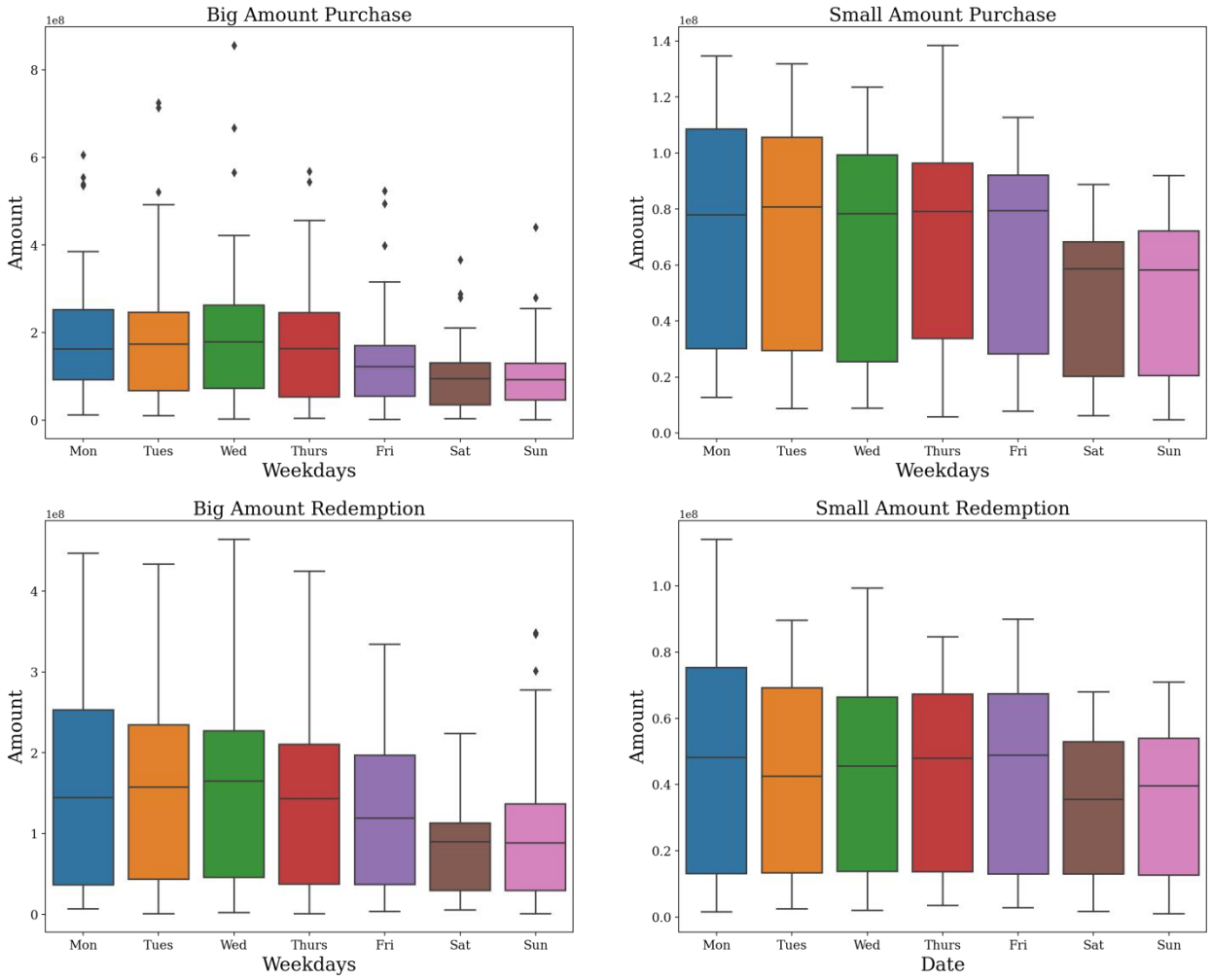

图 8 周内每日大额交易和小额交易比较

b. 申购和赎回异常交易行为可视化分析

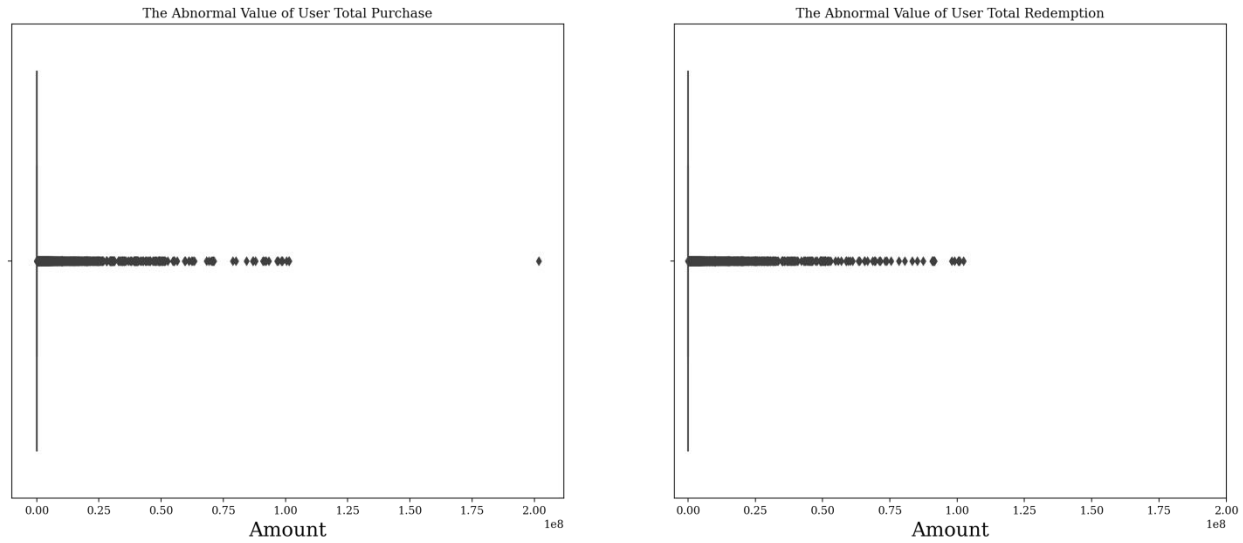

图 9 申购和赎回异常交易行为

图 9 展示了所有申购和赎回交易行为的箱型图。从图中可以看出，几乎所有申购和赎回交易的金额都集中于 10 万元以下，但申购中出现了一次“极端离群”交易行为，单日申购额超过 20 万。

c. 用户消费行为可视化分析

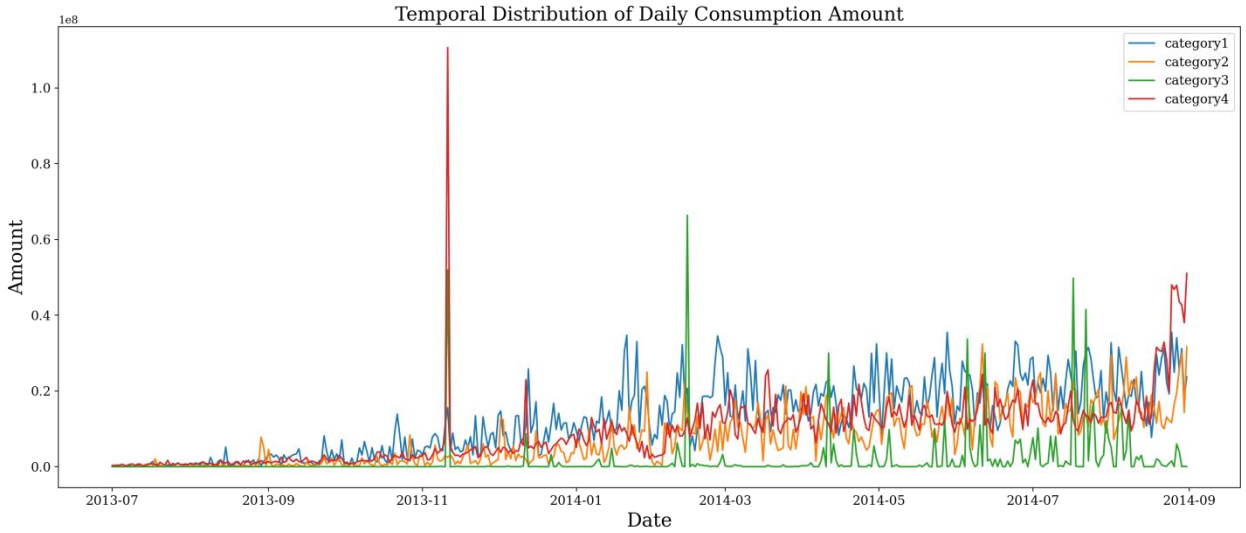

图 10 用户每日 4 类消费总数时序分布

图 10 展示了 2013 年 7 月到 2014 年 9 月用户每日 4 类消费总数的变化情况。从图中可以看出，2013 年 11 月出现了一次极高峰值，推测与“双十一”互联网消费节相关。除此之外，第一类消费整体高于另外三类消费。

（3）银行和支付宝利率可视化分析

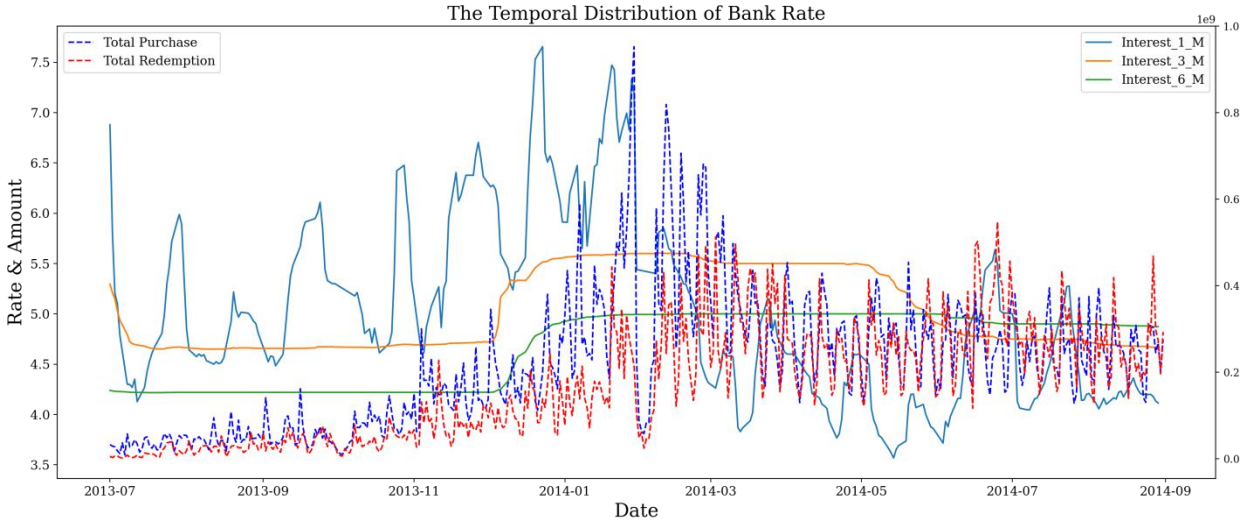

图 11 银行利率时序分布

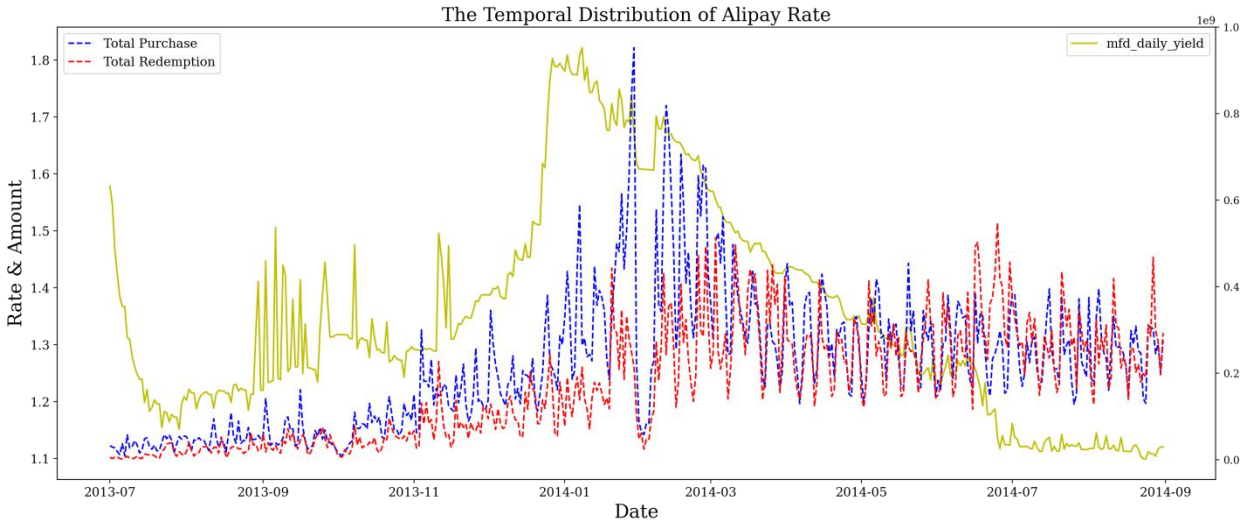

图 12 支付宝利率时序分布

图 11 和图 12 分别展示了 2013 年 7 月到 2014 年 9 月银行和支付宝的利率的变化。银行一个月利率波动较大，三个月和六个月利率与每日申购赎回总量出现了相似的趋势：2014 年的整体水平高于 2013 年。而支付宝的万分利息直观上与每日购买总量呈现出了一定相关性。

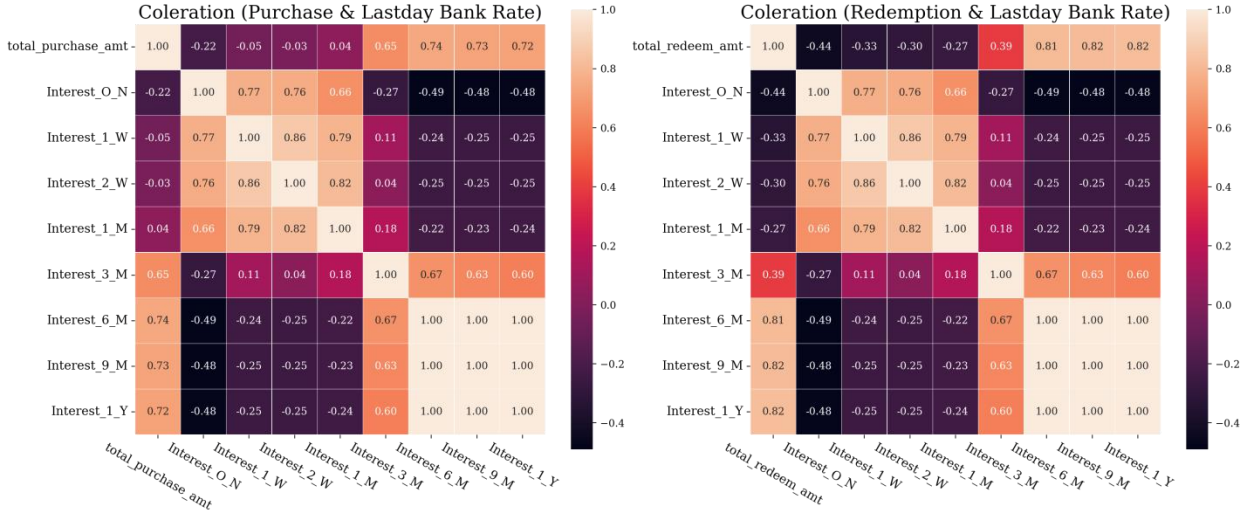

图 13 银行利率特征与申购和赎回的相关性矩阵

图 13 展示了银行利率特征与申购和赎回的相关性矩阵。由图中可以看出长期储蓄的利率与申购和每日赎回总量间的线性相关性均高于短期储蓄的利率与申购和每日赎回总量间的线性相关性。

但由图 10、11、12 可以看出，每日利率变化信息于 2014 年 9 月后未知，由于比赛官方要求 2014 年 8 月 31 号之后的公开数据不能用来预测，所以不考虑利用银行和支付宝相关信息构建特征。

四、 特征工程

(1) 特征构建

基于对数据集的探索性数据分析，我构建了以下三个维度共 46 项基本特征：

a. 基于日期的特征：根据每条数据的当天日期处理并提取了“是否是周末”、“是否是假期”等多项类别型特征，1 代表是，0 代表否。

b. 基于动态时序数据的特征：根据每条数据的当天日期到某特殊事件点的时间距离处理并提取了“距离放假还有多天”、“距离上班还有多少天”等多项数值型特征。为了防止部分数据的“距离放假还有多天”，“距离节假日还有多少天”和“距离节假日最后一天还有多少天”特征数值过大对建模预测产生影响，我对该这几项数据进行了特殊处理，若只要数值大于 5 则将该项数值固定为 10。

c. 基于动态时序数据的特征：以星期为周期统计当周申购和赎回总额的均值、中位数、最大值、最小值、标准差、偏度。对部分缺失值，采用每周申购或赎回数据的中位数进行填充。

(2) 特征总览

特征类别	特征名				
基于日期的特征	是否是周末	是否是假期	是否是特殊日子	是否是节假日的第一天	是否是节假日的最后一天
	是否是节假日后的上班第一天	是否不用上班	是否明天要上班	昨天上班了吗	是否是放假前一天
	是否是月初第一天	是否是月初第二天	是否是月初	是否是月中	是否是月末
	是否是每个月第 1 个周	是否是每个月第 2 个周	是否是每个月第 3 个周	是否是每个月第 4 个周	是否是星期一
	是否是星期二	是否是星期三	是否是星期四	是否是星期五	是否是星期六
	是否是星期日	共 26 项			
基于时间距离的特征	距离放假还有多天	距离上班还有多少天	距离节假日还有多少天	距离节假日最后一天还有多少天	距离月初第几天
	距离月的中心点有几天	距离星期的中心有几天	距离星期日有几天	共 8 项	
基于动态时序数据的特征	当周申购总量均值	当周申购总量中位数	当周申购总量最大值	当周申购总量最小值	当周申购总量标准差
	当周申购总量偏度	当周赎回总量均值	当周赎回总量中位数	当周赎回总量最大值	当周赎回总量最小值
	当周赎回总量标准差	当周赎回总量偏度	共 12 项		

表 2 特征总览表

(3) 特征选择

a. 相关性较低的特征的剔除

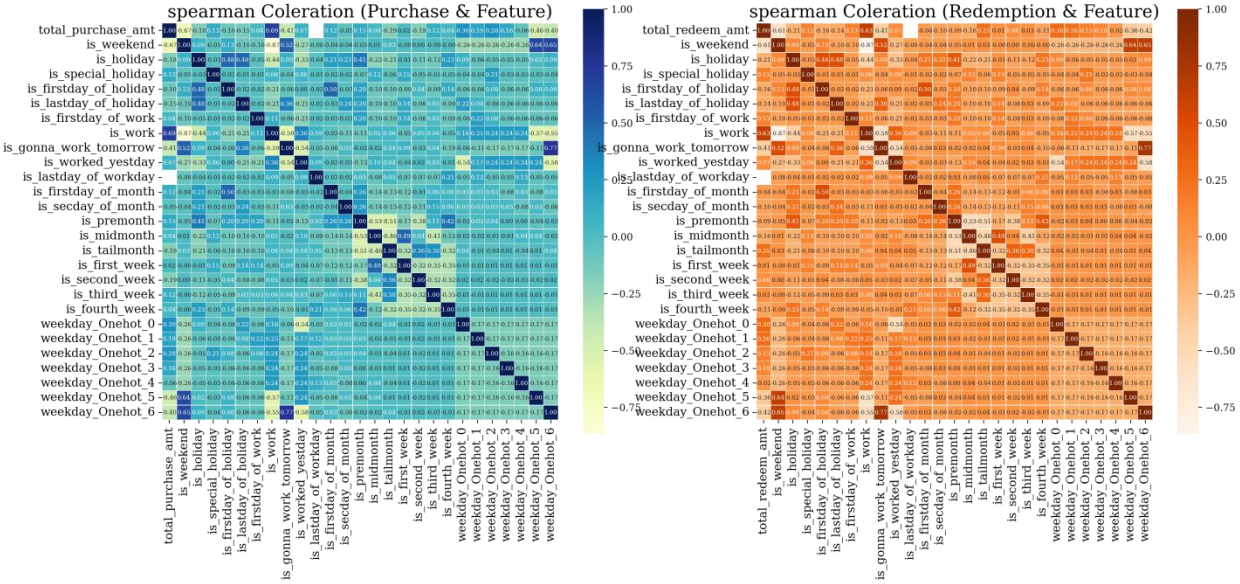

图 14 基于日期的特征与每日申购和赎回总量的斯皮尔曼相关性

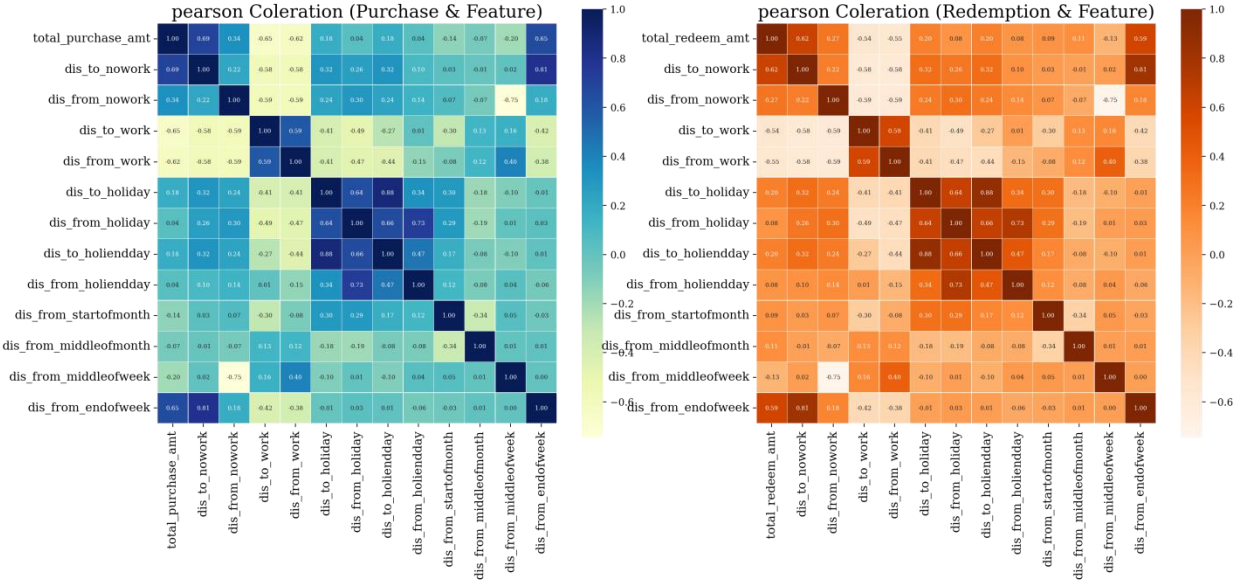

图 15 基于时间距离的特征与每日申购和赎回总量的皮尔逊相关性

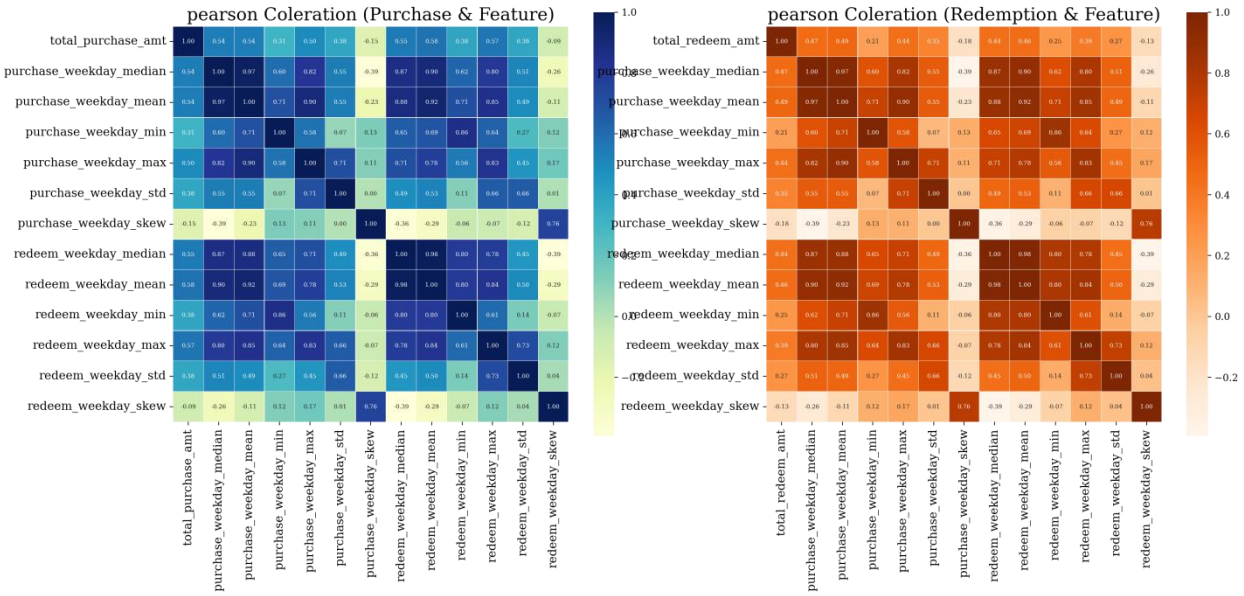

图 16 基于动态时序数据的特征与每日申购和赎回总量的皮尔逊相关性

图 14 展示了基于日期的特征与每日申购和赎回总量的斯皮尔曼相关性矩阵，图 15 和图 16 分别展示了基于时间距离的特征和基于动态时序数据的特征与每日申购和赎回总量的皮尔逊相关性。与标签之间线性相关性较低的特征对于建模预测没有帮助，我将所有与标签相关性小于 0.1 的特征直接剔除。由于将对每日申购总量和赎回总量分别建模，因此对申购和赎回的也分别进行特征选择，被剔除的相关性较低特征如表 3：

	相关性小于 0.1 的特征
申购特征集合	'is_firstday_of_work', 'is_worked_yestday', 'is_fourth_week', 'dis_from_middleofmonth', 'is_midmonth', 'weekday_Onehot_4', 'is_first_week', 'dis_from_holiday', 'redeem_weekday_skew', 'dis_from_holiendday', 'is_secday_of_month'
赎回特征集合	'is_second_week', 'is_third_week', 'dis_from_startofmonth', 'is_worked_yestday', 'weekday_Onehot_4', 'is_first_week', 'dis_from_holiday', 'is_premonth', 'is_firstday_of_month', 'dis_from_holiendday'

表 3 被剔除的相关性较低特征

b. 对复共线性特征的处理

复共线性是指在回归分析中当自变量的个数很多，且相互之间相关很高(大于等于 0.8)，自变量之间存在近似的线性关系，会导致样本资料估计的回归系数的精度显著下降。因此，我们需要将两个具有复共线性的特征融合成一个特征，并剔除冗余的特征。最终对复共线性特征如表 4 所示：

	复共线性特征
申购特征集合	1. 'dis_from_nowork' 和 'dis_from_middleofweek'=> 结合 'dis_from_nowork%%%%dis_from_middleofweek' 2. 'dis_to_holiendday'和'is_holiday'=>结合为'dis_to_holiendday%%%%is_holiday'
赎回特征集合	1. 'dis_from_nowork' 和 'dis_from_middleofweek'=> 结合 'dis_from_nowork%%%%dis_from_middleofweek'

	2. 'dis_from_middleofmonth' 和 'is_midmonth'=> 结合为 'dis_from_middleofmonth%%%%is_midmonth'

表 4 对复共线性特征的处理

c. 基于变量重要性的特征筛选

① 基于排序重要性的特征选择

对某一特征进行随机排序应当会降低预测的准确率，这是因为产生的数据不再对应于现实世界中的任何特征意义。如果随机打乱的那一列模型预测对其依赖程度很高，那么模型准确率的衰减程度就会更大, 则排序重要性更高。

我将 2014 年 3 月 31 日到 2014 年 7 月 31 日的数据划分为训练集, 2014 年 8 月 1 日至 31 日的数据划分为验证集, 利用线性回归模型分别对每日申购总量和每日赎回总量进行建模预测，并计算所有的特征排序重要性，分别记录前 20 重要的特征，各排序重要性如下图 17 所示：

Weight	Feature
2.7207 ± 1.8471	dis_from_nowork%%%%dis_from_middleofweek
2.2538 ± 1.0352	weekday_Onehot_3
2.2143 ± 0.4414	dis_from_nowork
1.2732 ± 0.4930	purchase_weekday_std
0.6929 ± 0.4346	purchase_weekday_min
0.4472 ± 0.3079	purchase_weekday_max
0.3396 ± 0.1612	is_tailmonth
0.3192 ± 0.2264	redeem_weekday_mean
0.2965 ± 0.2818	is_third_week
0.2348 ± 0.2433	redeem_weekday_std
0.1559 ± 0.1703	is_firstday_of_month
0.0737 ± 0.1973	is_work
0.0574 ± 0.0739	dis_from_startofmonth
0.0573 ± 0.0647	weekday_Onehot_0
0.0470 ± 0.0366	purchase_weekday_skew
0.0176 ± 0.0175	is_gonna_work_tomorrow
0.0162 ± 0.0776	weekday_Onehot_2
0.0125 ± 0.0530	weekday_Onehot_5
0.0000 ± 0.0000	is_special_holiday
0.0000 ± 0.0000	is_lastday_of_workday
... 8 more ...	

Weight	Feature
13.1299 ± 4.3049	dis_from_nowork%%%%dis_from_middleofweek
7.7271 ± 4.1117	weekday_Onehot_3
3.2140 ± 2.9837	is_work
1.8521 ± 0.2437	weekday_Onehot_2
1.2222 ± 0.5126	dis_from_nowork
0.8594 ± 0.3924	redeem_weekday_min
0.4818 ± 0.3458	purchase_weekday_mean
0.3063 ± 0.2532	is_gonna_work_tomorrow
0.1603 ± 0.2184	dis_from_endofweek
0.1100 ± 0.0691	weekday_Onehot_5
0.1090 ± 0.1647	purchase_weekday_skew
0.0972 ± 0.1747	dis_from_middleofmonth%%%%is_midmonth
0.0828 ± 0.2071	weekday_Onehot_0
0.0600 ± 0.1255	is_tailmonth
0.0243 ± 0.0649	purchase_weekday_std
0.0149 ± 0.0350	dis_from_middleofmonth
0.0088 ± 0.0538	total_purchase_amt
-0.0000 ± 0.0000	is_lastday_of_workday
-0.0000 ± 0.0000	is_special_holiday
-0.0000 ± 0.0000	is_firstday_of_holiday
... 8 more ...	

图 17 对每日申购量（左）和每日赎回量（右）预测的线性回归模型中各特征排序重要性

② 基于 SHAP value 的特征选择

SHAP 是 Python 开发的一个"模型解释"包，可以解释任何机器学习模型的输出。其名称来源于 SHapley Additive exPlanation，在合作博弈论的启发下 SHAP 构建一个加性的解释模型，所有的特征都视为“贡献者”。对于每个预测样本，模型都产生一个预测值，SHAP value 就是该样本中每个特征所分配到的数值, 其计算方法此处不做赘述。

同样将 2014 年 3 月 31 日到 2014 年 7 月 31 日的数据划分为训练集, 2014 年 8 月 1 日至 31 日的数据划分为验证集, 利用 XGBoost 回归模型分别对每日申购总量和每日赎回总量进行建模预测，并计算每个特征的 SHAP value，各 SHAP value 如下图 18 和图 19 所示：

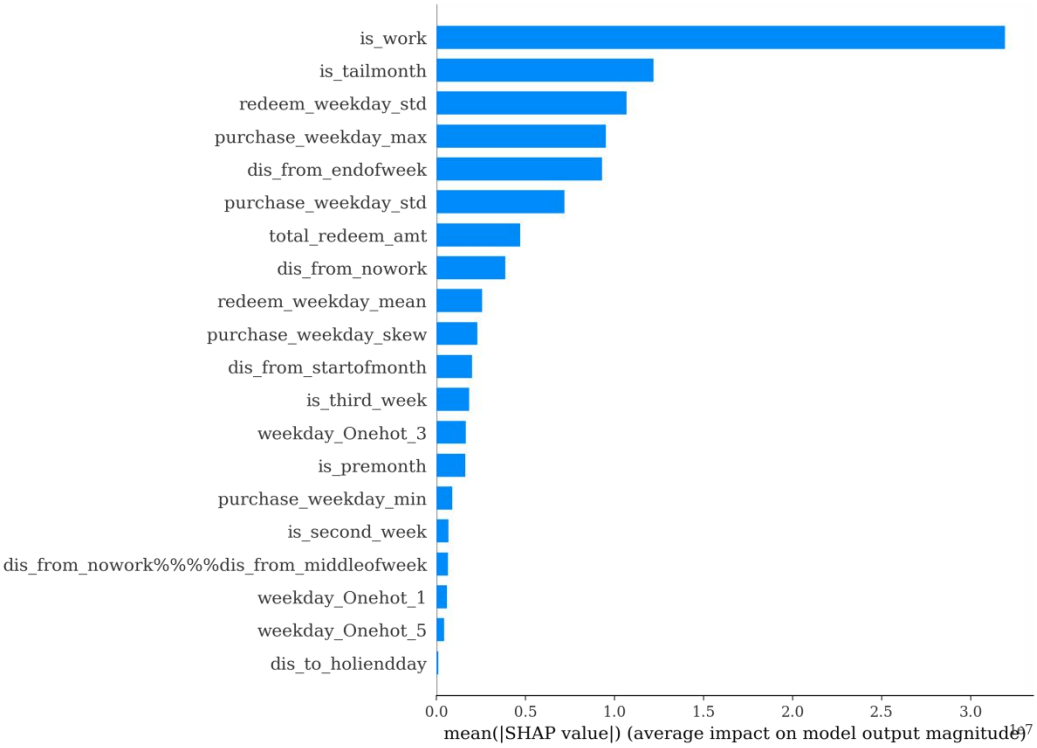

图 18 XGBoost 对每日申购总量预测的 SHAP value

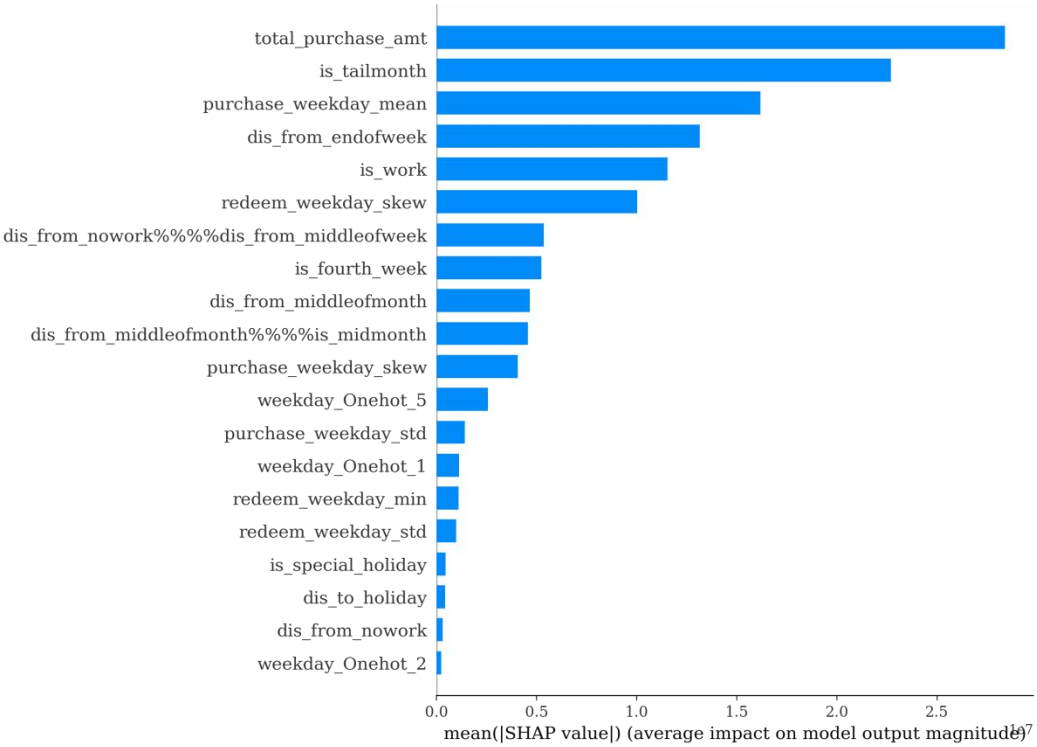

图 19 XGBoost 对每日赎回总量预测的 SHAP value

最后将①和②中选中的特征集合取交集选出最终的优胜特征，最终优胜特征如下表 5 所示：

	最终优胜特征
申购建模优胜特征	'redeem_weekday_std', 'weekday_Onehot_5', 'dis_from_nowork', 'total_redeem_amt', 'is_premonth', 'purchase_weekday_min', 'weekday_Onehot_2', 'weekday_Onehot_0', 'dis_from_endofweek', 'weekday_Onehot_1', 'purchase_weekday_max', 'purchase_weekday_std', 'purchase_weekday_skew', 'is_firstday_of_month', 'redeem_weekday_mean', 'is_work', 'is_tailmonth', 'is_gonna_work_tomorrow', 'is_second_week', 'dis_to_holiendday'
赎回建模优胜特征	'redeem_weekday_std', 'weekday_Onehot_3', 'dis_from_nowork', 'total_purchase_amt', 'weekday_Onehot_2', 'weekday_Onehot_0', 'dis_from_endofweek', 'weekday_Onehot_1', 'purchase_weekday_std', 'dis_to_holiday', 'is_work', 'is_tailmonth', 'purchase_weekday_mean', 'is_lastday_of_holiday', 'is_firstday_of_work', 'dis_from_middleofmonth', 'redeem_weekday_skew', 'is_special_holiday', 'redeem_weekday_min', 'is_secday_of_month'

表 5 最终优胜特征集

五、 模型评估

（1）基线模型

为了实现对未来 30 天内每日的资金流入流出情况的预测，我选用了以下 6 种数理统计或机器学习模型：

a. 线性回归模型：利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法

b. 决策树回归模型：是在已知各种情况发生概率的基础上，通过构成决策树来求取净现值的期望值大于等于零的概率，评价项目风险，判断其可行性的决策分析方法，是直观运用概率分析的一种图解法。

c. 随机森林回归模型：随机森林指的是利用多棵决策树树对样本进行训练并预测的一种模型。

d. 梯度提升回归模型：梯度提升（Gradient Boosting）是一种提升方法，也是一种常用于回归和分类问题的集成学习算法和机器学习技术，以弱预测模型（通常是决策树）集合的形式产生预测模型。

e. 多层感知机回归模型：多层感知器是一种前馈人工神经网络模型，其将输入的多个数据集映射到单一的输出的数据集上。

f. XGBoost 回归模型：XGBoost 是一种基于梯度提升的算法。其基本思想和梯度提升相同，但是做了一些优化，比如二阶导数使损失函数更精准；正则项避免树过拟合；Block 存储可以并行计算等。

（2）实验设置

我将 2014 年 3 月 31 日到 2014 年 8 月 31 日的数据划分为训练集进行模型的拟合和训练，预测申购的模型采用表 5 中申购建模优胜特征，预测赎回的模型采用表 5 中赎回建模优胜特征，所有数值型特征均经过均值归一化处理。所有模型超参数都采用 Python sklearn 库或者 xgboost 库中的模型默认参数，未进行特殊调参。

（3）线上预测结果

最终 XGBoost 回归模型实现了最高的线上预测得分（132.8909 分，如图 20 所示），该得分由天池在线平台给出，各类模型的线上预测得分如表 6 所示。

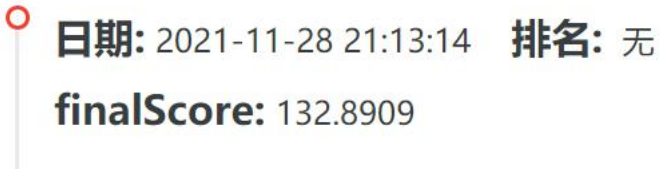

图 20 最终线上预测得分

模型	finalScore
LinearRegression	123.7332
DecisionTreeRegressor	104.1607
RandomForestRegressor	118.6584
GradientBoostingRegressor	119.0291
MLPRegressor	119.3555
XGBRegressor	**132.8909**

表 6 线上预测结果

六、 总结

在这个项目中，我通过对余额宝用户的历史申购赎回数据、用户个人信息数据和银行利率波动等业务相关数据的把握，实现对未来 30 天内每日的资金流入流出情况的预测。

我首先对时序数据、用户交易和消费行为数据和银行及支付宝利率数据的可视化分析。基于先前的分析，我构建了基于日期的特征，基于时间距离和基于动态时序数据三个维度共 46 项基本特征。通过去除相关性较低的特征选择，处理复共线性特征以及基于多种特征重要性准则的特征筛选，我确定了最终采用的申购和赎回建模优胜特征集合。

我选择了 6 种数理统计或机器学习模型进行训练和拟合，最终 XGBoost 回归模型实现了最高的线上预测得分：132.8909 分。

由于时间限制，目前所达到的回归精度还能得到进一步提升，在未来，我可能通过以下几个方向进一步提升预测效果：

（1） 降低建模细粒度。当前只对申购和赎回两项指标实现了预测，但实际上，建模细粒度可以继续降低，例如：今日申购总量=今日支付宝余额购买+今日银行卡购买+今日收益购买，可以将对今日申购总量的预测细粒化为对今日支付宝余额购买、今日银行卡购买和今日收益购买的预测。

（2） 对 Xgboost 模型调参。为公平比较，所有模型超参数都采用 Python sklearn 库或者 xgboost 库中的模型默认参数，未进行特殊调参，对 Xgboost 进一步调参可能提升模型的精度。

（3） 进一步特征构造：继续挖掘和融合新的特征，有效的利用更多数据信息。