基于深度学习算法的资金流向预测模型

项目概要介绍

WTH 团队

2019.03

一、项目介绍与目标

1.1 项目背景

随着科学技术的发展，互联网金融正在逐步成为金融体系中的终于组成部分。随着可随存随取的基金的普及和壮大，每天都会涉及极为大量的资金流入和流出，其中首当其冲的就是阿里巴巴旗下蚂蚁金服运作的余额宝。

蚂蚁金服是一家定位于惠普金融服务的科技企业，起步于 2004 年成立的支付宝。截至 2019 年末，支付宝累计交易用户突破 6 亿户。面对庞大的用户群体，资金头寸管理的压力非常大。所以，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。如果能相对精确地预测出每日的资金流动情况，就可以在保证安全的情况下，对可用的资金仅用规划，取得最大收益。由此看来，有效准确地对资金流入流出进行预测，将具有广阔的市场前景和巨大的市场价值。

1.2 项目简介

本项目的核心是通过对余额宝用户的申购赎回行为数据和用户信息的把握，对未来一个月内每日的资金总流入流出情况进行预测。通过 ARIMA、LSTM 等数学模型进行解决并分析比较，得出最合适的方法，获得较为精准的预测结果。

我们将通过三万余额宝 用户在 13 个月内（2013 年 7 月——2014 年 8 月）的脱敏加密的完整行为数据（购买/赎回）和简单的用户信息，对未来（2014 年 9 月）每一天的资金申购和赎回总额进行预测，预测结果精确到分。

1.3 项目目标

通过合理的方法建立预测模型，基于用户信息和用户申购赎回行为数据，对未来的资金总流入总厨情况进行精准预测。

1.4 项目解决思路

1.5 项目创新点

- 结合智能算法。运用 ARIMA、LSTM、LM 多种深度学习算法，科学性强，预测趋势准确。
- 构成合理，流程完整。在构建预测模型前进行了较为全面的数据分析和预处理，包括关联性分析和数据可视化，为后续项目的合理展开提供了保障。
- 多种模型结合，弥补不同模型间的缺陷。
- 可操作性强，并可不断进行优化，进步空间大

1.6 项目团队

项目团队共有软件学院五名本科生组成。

队长矫丽瑶 软件学院软件工程本科大三在读生。本人曾参与全国大学生数学建模比赛获省一等奖，参与有关 GAN 的相关项目，掌握一些相关的变体模型。在本项目中主要负责项目的策划，参与

技术的实现，把握项目进度。

队员刘悦然，软件学院软件工程本科大三在读生。在 2018 高教社杯全国大学生数学建模竞赛中获福建省一等奖，曾系统学习卷积神经网络相关内容，为厦门大学大创项目卷积神经网络在语音转换方面的应用的负责人。在此次比赛项目参与负责初期模型的选择与构建，以及完成对项目的市场分析调研等相关工作。

队员牛晓彤，厦门大学软件学院软件工程本科大三在读生。本人曾参与美国数学建模比赛，主要负责模型的分析和 Matlab 的编程实现。在本项目中，本人主要负责模型的实现与分析。

成员苏天宇，厦门大学软件学院软件工程大三本科生。本人曾参与大学翻转课堂开发项目，负责前端开发和优化。在本项目中，本人主要负责赛题的时间序列预测方法的研究和技术实现。

组员王世奇 厦门大学软件学院软件工程本科。本人曾参与 GAN 神经网络的学习项目，研究过神经网络的基础原理和一些 GAN 改进模型的大体框架。在本项目中，本人作为项目组成员主要进行数据分析，编写代码，参数调优等工作。

二、 产品与服务

2.1 产品介绍

为减轻企业对资金管理的压力，有效的对可用资金进行规划，本项目根据用户的行为数据和用户信息，利用 ARIMA、LSTM 等数学模型对之后每一天的资金申购和赎回总额进行预测，以判断资金流入流出的情况。

2.1.1 数据准备

总共拥有用户信息表、用户申购赎回数据表、收益率表和银行间同业拆放利率四张表。

1. 用户信息表

用户信息表提供了用户的基本信息。在原始数据的基础上处理后，主要包括了用户的性别、城市和星座

列名	类型	含义	示例
user_id	bigint	用户 ID	1234
Sex	bigint	用户性别 （1:男，0:女 ）	0
City	bigint	所在城市	6081949
constellation	string	星座	射手座

表格 2-1 用户信息表

2. 用户申购赎回表

用户申购赎回表包含 20130701 至 20140831 的申购和赎回信息以及所有子类目信息，数据已经过脱敏处理（基本保持了原数据趋势）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。

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

category2	bigint	今日类目 2 消费总额	0
category3	bigint	今日类目 3 消费总额	0
category4	bigint	今日类目 4 消费总额	0

表格 2-2 用户申购赎回表

注 1 :上述的数据都是经过脱敏处理的，收益为重新计算得到的，计算方法按照简化后的计算方式处理，具体计算方式在下节余额宝收益计算方式中描述。

注 2 :脱敏后的数据保证了今日余额 = 昨日余额 + 今日申购 - 今日赎回，不 会出现负值。

3. 收益率表

收益率表为余额宝在 14 个月内的收益率表

列名	类型	含义	示例
mfd_date	string	日期	20140102
mfd_daily_yield	double	万份收益，即一万块钱的收益	1.5787
mfd_7daily_yield	double	七日年化收益率（%）	6.307

表格 2-3 收益率表

4.银行间同业拆放利率表

银行间拆借利率表是 14 个月期间银行之间的拆借利率（皆化为年化利率）

列名	类型	含义	示例
mfd_date	String	日期	20140102
Interest_O_N	Double	隔夜利率（%）	2.8
Interest_1_W	Double	1 周利率（%）	4.25

Interest_2_W	Double	2 周利率（%）	4.9
Interest_1_M	Double	1 个月利率（%）	5.04
Interest_3_M	Double	3 个月利率（%）	4.91
Interest_6_M	Double	6 个月利率（%）	4.79
Interest_9_M	Double	9 个月利率（%）	4.76
Interest_1_Y	Double	1 年利率（%）	4.78

表格 2-4 银行间同业拆放利率表

2.1.2 收益计算方式

收益计算方式主要基于实际余额宝收益计算方法，单进行但进行了一定的简化，计算简化的地方如下：

1）收益计算的时间不再是会计日，而是自然日，以 0 点为分隔，如果 是 0 点之前转入或者转出的金额算作昨天的，如果是 0 点以后转入或者转出的金 额则算作今天的。

2）收益的显示时间，即实际将第一份收益打入用户账户的时间为如下表 格，以周一转入周三显示为例，如果用户在周一存入 10000 元，即 1000000 分， 那么这笔金额是周一确认，周二是开始产生收益，用户的余额还是 10000 元，在周三将周二产生的收益打入到用户的账户中，此时用户的账户中显示的是 10001.1 元，即 1000110 分。其他时间的计算按照表格中的时间来计算得 到。

简化后的余额宝收益计算表

转入时间	首次显示收益时间
周一	周三
周二	周四
周三	周五
周四	周六
周五	下周二
周六	下周三
周天	下周三

表格 2-5 余额宝收益计算表

2.2 产品特色

2.2.1 独创性

目前市场上还未存在同类型企业项目，市场相对宽松，有较大的发展空间

2.2.2 训练快捷

预测模型利用 ARIMA、LSTM、HMM 等深度学习模型，所需训练数据量小，预测速度快。

2.2.3. 预测单位精准

模型预测趋势较为精准，可精确到每天。

2.2.4. 可提供个性化服务

项目可提供个性化服务，综合考虑多种因素，根据用户需求来完成不同程度的预测。

2.3 主要功能介绍

本项目利用 ARIMA、LSTM 等深度学习模型，能够通过以往用户的行为操作和相关信息，来较为准确的预测出未来时间段内资金的流出和流出情况，使得金融公司和电商企业等相关产品用户可以利用该趋势调动资金，进而更好的进行发展。

三、 技术路线与实现方案

3.1 评估指标

采用积分式的计算方法。每天的误差选用相对误差来计算，然后根据用户预测申购和赎回的相对误差，通过得分函数映射得到一个每天预测结果的得分，将 30 天内的得分汇总，然后结合实际业务的倾向，对申购赎回总量预测的得分情况进行加权求和，得到最总评分。

1. 计算所有用户在测试集上每天的申购和赎回总额与实际情况总额的误差。

每日申购相对误差(真实值z_i，预测值为$\hat{z}_i$)：

$$\text{Purchase}_i = \frac{|z_i - \hat{z}_i|}{z_i}$$

每日赎回相对误差(真实值y_i，预测值为$\hat{y}_i$)：

$$\text{Redeem}_i = \frac{|y_i - \hat{y}_i|}{y_i}$$

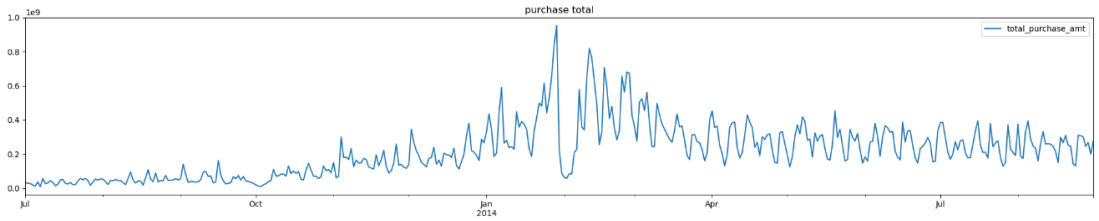
2. 申购预测得分 Purchasei 相关，赎回预测得分与 Redeemi 相关 ，误 差与得分之间的计算公式不公布，但保证该计算公式为单调递 减的， 即误差越小，得分越高，误差与大，得分越低。当第 i 天的申购误差 Purchasei=0 ，这一天的得分为 10 分;当 Purchasei ＞ 0.3 ，其得分为 0 。

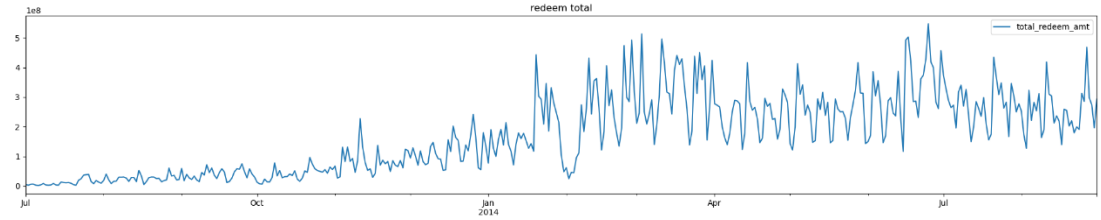
3. 最后公布总积分 = 申购预测得分 *45%+ 赎回预测得分 *55%

3.2 数据分析与预处理

3.2.1 总体分析图

1. 时间——申购赎回每日总额时序图



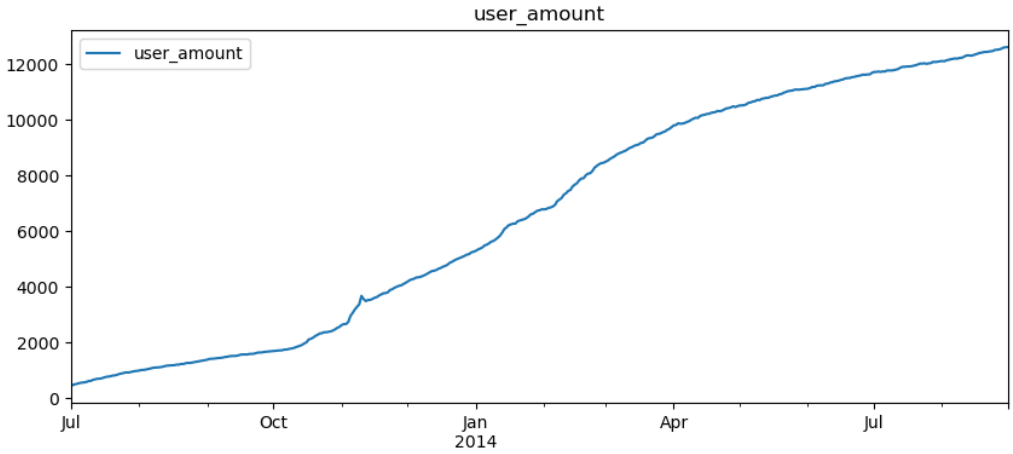


由图可知，本项目目标时序均不平稳，具有上升趋势明显，异常值多，方差波动极其复杂的特点，需要进一步分析其潜在的影响因素，将其分解为更具解释性的子部分。

3.2.2 基础信息图

1. 时间——人数图

由图可知，本项目的日用户量前中期满足 logistic 人口增长模型的特征，而后期具有极佳的线性度，具有非常好的拟合特性，而其增长趋势与申购赎回总额的增长趋势相似，可以用于拟合目标时序的趋势性。



1. 星座——人数图

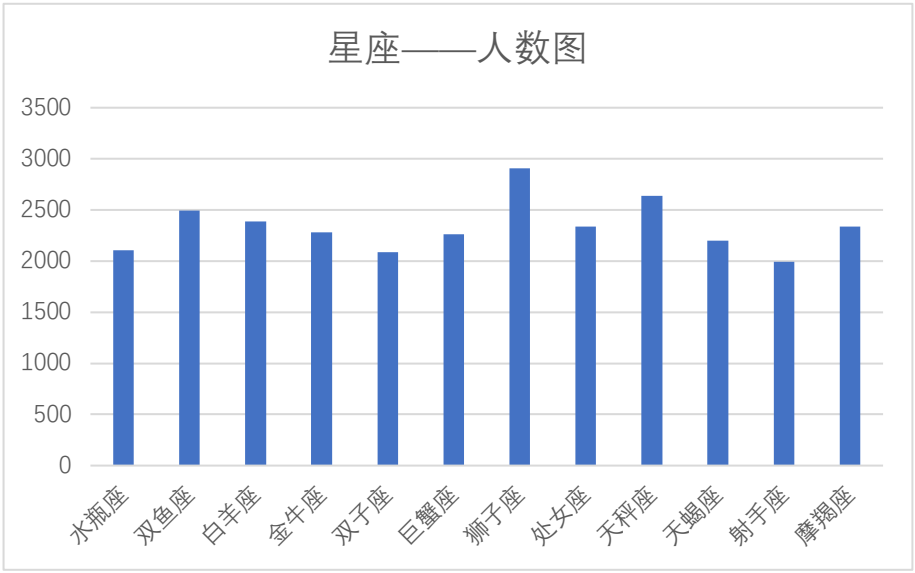


图 3-0-1 星座-人数图

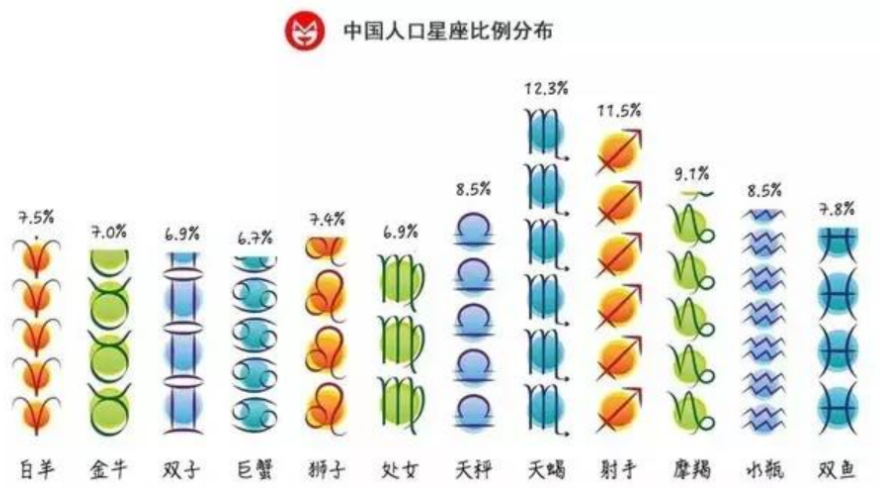


图 3-0-2 中国人口星座比例分布

根据查找到的中国人口星座比例分析图与本项目数据相比，可知本项目星座_

人数分布与大环境统计结果并不完全一致，但本项目各星座分布仍属均衡（2000-2500），无极限值出现，且无法准确考虑各星座的消费存储差异，则在本项目的分析预测中可不予考虑。

2. 城市——人数图

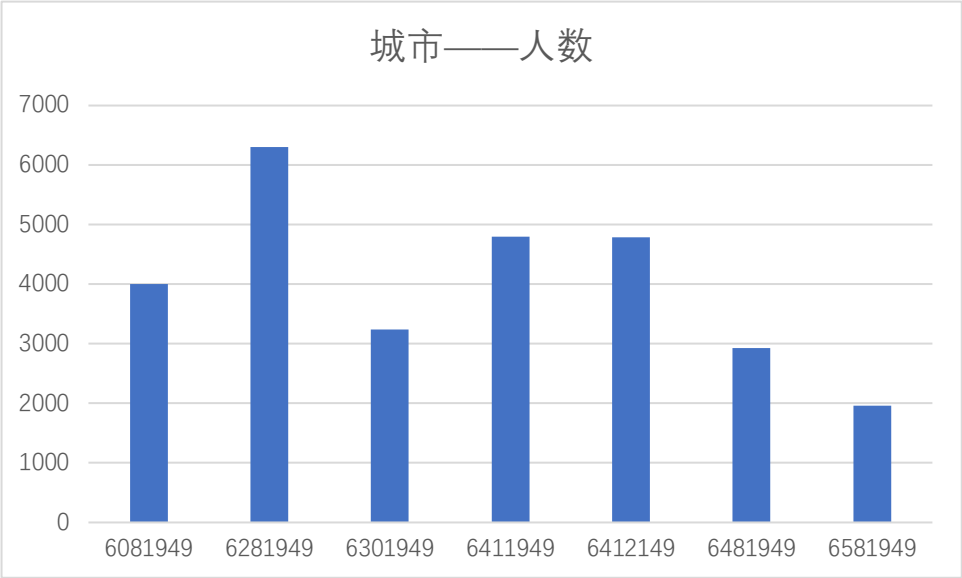


图 3-0-3 城市-人数图

由图可知，不同城市的人口分布还是有些差异，但项目说明中仅出示城市代号编码，并未给出具体城市信息，对于城市的消费习惯不可仅从人数区分，可以考虑各城市购买消费均值来分类城市

3. 时间——余额宝利率

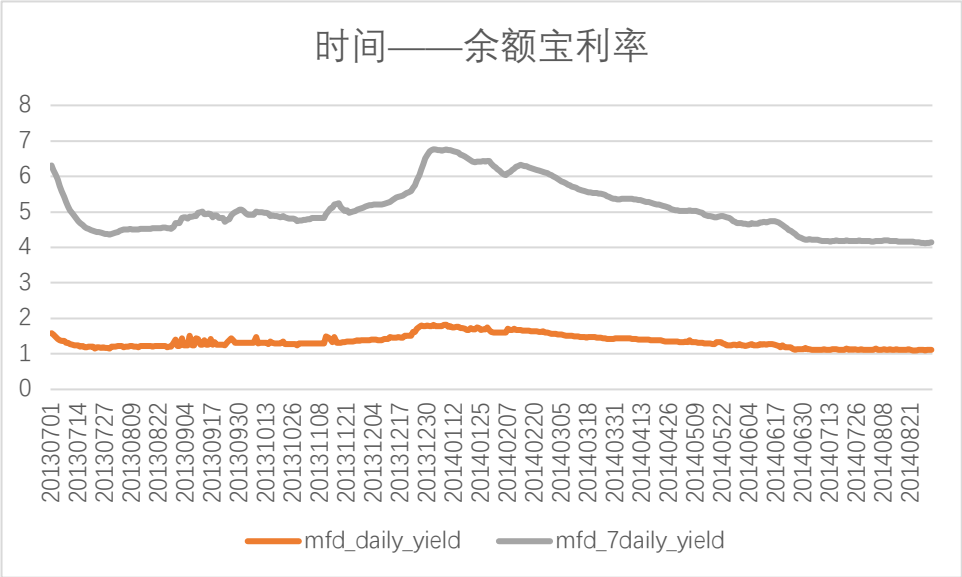


图 3-0-4 时间-余额宝利率图

此图可知，余额宝利润在这 14 个月中有较大变动，由于预测月份为第 15 个月，由于利率在前 8 个月利率波动较大且距离预测月份较远（影响较小），则可考虑根据后 6 个月比较平稳的利率来准确预测第 15 个月的数据

4. 时间——银行利率

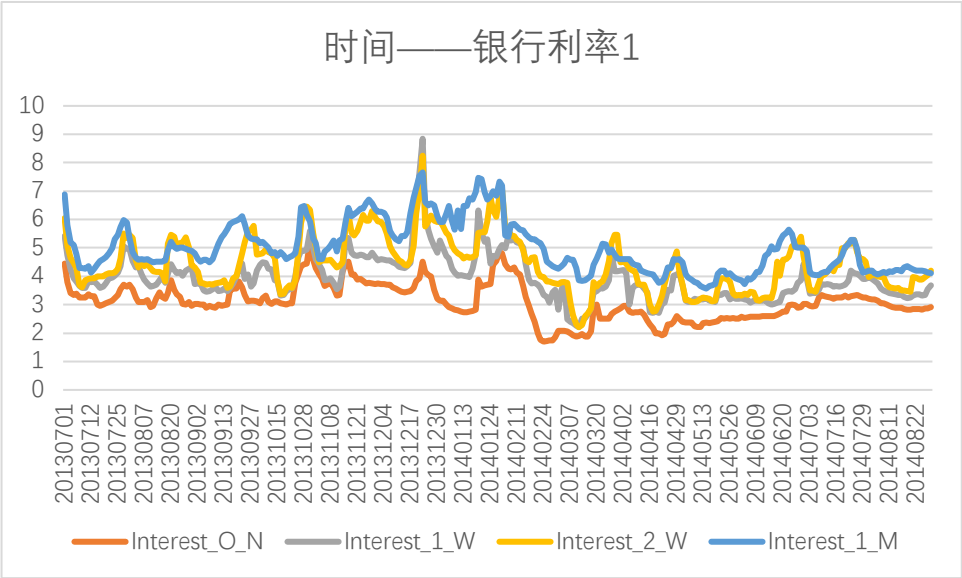


图 3-0-5 时间-银行利率图 1

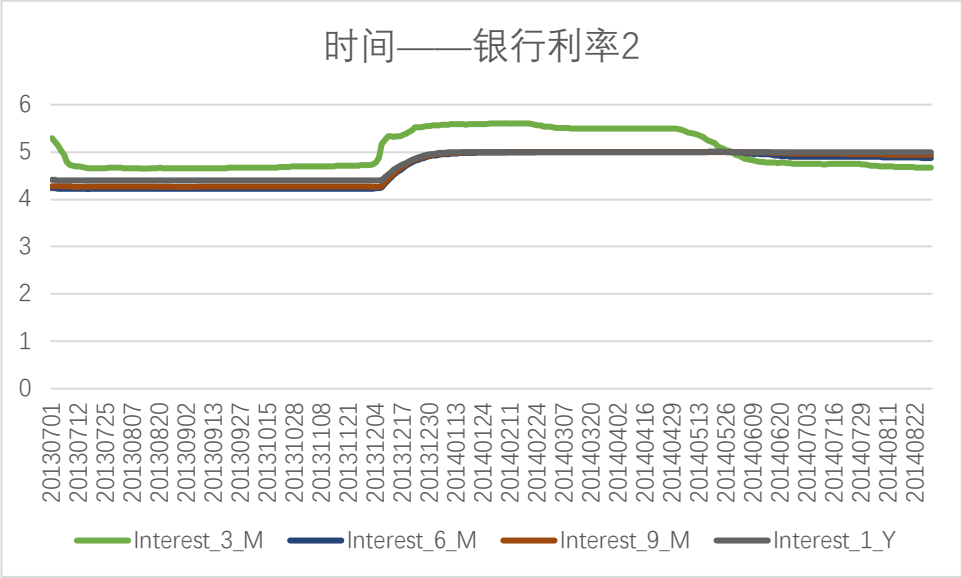


图 3-0-6 时间-银行利率图 2

由图像可以得出，银行的的短期利率（包括隔夜、一周、两周、一个月）几乎呈现相同的变化趋势，相对来说利率波动较大且频繁，不是很稳定；银行的长期利率（包括三个月、六个月、九个月、一年）几乎呈现相同的变化趋势，相对来说波动较小且稳定。可据此将利率的变化曲线简单划分成两类

3.2.2 关联性分析

由于目前存在众多可能会对用户申购赎回行为及其金额产生影响的因素，如用户的性别、星座、城市、利率等等。为了进一步确定挖掘已存在的因素之间的关联性和相关性，合理地进行后续数据分析，项目采用 FP-Growth 算法进行关联分析。

因为不便直接挖掘各因素和具体金额的之间的关联，所以项目根据用户的余额和行为数据将用户划分成‘rich‘和‘poor’两个等级。通过分析用户因素与用户等级的

关系，进而得到可能会对金额产生影响的关键因素。

1. 频繁模式挖掘

将 user_rpofile_table 和 user_balance_table 数据表连接，并以 500000 为分界线，用户的，以此产生新的数据集 user_balance_classification，并将其转换为 FP-Tree 所需要的格式，当支持度 sup 分别为 0.005 和 0.001 时，所得的频繁模式如下：

pattern（sup=0.005）.csv | patter(sup=0.001).csv

```
pattern,counts
poor-6581949,1737
1-6581949,1036
poor-1-6581949,911
AR-6581949,159
poor-1-AR-6581949,80
LI-6581949,189
poor-1-LI-6581949,87
CA-6581949,151
poor-1-CA-6581949,68
CP-6581949,143
poor-1-CP-6581949,65
PI-6581949,179
poor-1-PI-6581949,72
AQ-6581949,181
poor-1-AQ-6581949,89
GE-6581949,175
poor-1-GE-6581949,81
Vi-6581949,159
poor-1-Vi-6581949,73
Le-6581949,178
poor-1-Le-6581949,89
SC-6581949,156
poor-1-SC-6581949,75
SG-6581949,151
poor-1-SG-6581949,68
TA-6581949,144
poor-1-TA-6581949,64
0-6581949,929
poor-0-6581949,826
poor-0-PI-6581949,78
poor-0-CP-6581949,64
poor-0-CA-6581949,57
poor-0-TA-6581949,63
```

图 3-0-7 sup=0.001 时的频繁模式

pattern（sup=0.005）.csv × patter(sup=0.001).csv ×

```
poor-1-SG,841
0-SG,963
6411949-SG,355
6412149-SG,338
poor-1-6412149-SG,155
6281949-SG,467
poor-1-6281949-SG,186
poor-1-6411949-SG,172
6481949-SG,196
6081949-SG,287
6301949-SG,198
poor-0-SG,791
poor-0-6281949-SG,167
rich-SG,360
poor-GE,1718
1-GE,1072
poor-1-GE,889
6412149-GE,364
poor-1-6412149-GE,169
6281949-GE,426
poor-1-6281949-GE,163
6411949-GE,349
poor-1-6411949-GE,166
6481949-GE,229
6081949-GE,319
6301949-GE,226
0-GE,1016
poor-0-GE,829
poor-0-6281949-GE,164
rich-GE,370
poor-AQ,1739
1-AQ,1102
poor-1-AQ,914
```

图 3-0-8 sup=0.005 时的频繁模式

2. 关联规则挖掘

根据已得的频繁模式，调整最小置信度 minconfidence 的值，得到的规则结果部分如下

rules(sup=0.001 mincon=0.6).csv

```
rule,confidence,support
6581949->poor,0.8839694656488549,1737
1-6411949->poor,0.8800922367409685,2290
1-6581949->poor,0.8793436293436293,911
6411949->poor,0.8765354986466791,4210
6481949->poor,0.8753840901331512,2564
1-6481949->poor,0.8659722222222223,1247
6412149->poor,0.8482571488207055,4064
1-AR->poor,0.8456486042692939,1030
1-TA->poor,0.8401332223147377,1009
AR->poor,0.8361960620025136,1996
1-Le->poor,0.8343393695506371,1244
SC->poor,0.8337874659400545,1836
PI->poor,0.8317981577893472,2077
1-PI->poor,0.8311588641596316,1083
CA->poor,0.8295805739514349,1879
1-AQ->poor,0.8294010889292196,914
1-GE->poor,0.8292910447761194,889
1-SC->poor,0.8280977312390925,949
1->poor,0.8278241667241736,11996
1-CA->poor,0.8275261324041812,950
1-Vi->poor,0.8261603375527427,979
AQ->poor,0.8249525616698292,1739
Vi->poor,0.824486301369863,1926
GE->poor,0.8227969348659003,1718
1-CP->poor,0.8223308883455582,1009
TA->poor,0.8197368421052632,1869
SG->poor,0.8192771084337349,1632
0->poor,0.8187117243414742,11096
Le->poor,0.8164948453608247,2376
0-GE->poor,0.8159448818897638,829
CP->poor,0.8129280821917808,1899
LI->poor,0.8121212121212121,2144
```

图 0-9 sup=0.001 mincon=0.6 的关联规则

rules(sup=0.005. mincon=0.75).csv

```
rule,confidence,support
0-6581949->poor,0.8891280947255114,826
0-6481949->poor,0.8844862323707186,1317
6581949->poor,0.8839694656488549,1737
1-6411949->poor,0.8800922367409685,2290
1-6581949->poor,0.8793436293436293,911
6411949->poor,0.8765354986466791,4210
6481949->poor,0.8753840901331512,2564
1-6481949->poor,0.8659722222222223,1247
6412149->poor,0.8482571488207055,4064
1-AR->poor,0.8456486042692939,1030
AR->poor,0.8361960620025136,1996
1-Le->poor,0.8343393695506371,1244
SC->poor,0.8337874659400545,1836
PI->poor,0.8317981577893472,2077
1-PI->poor,0.8311588641596316,1083
CA->poor,0.8295805739514349,1879
1->poor,0.8278241667241736,11996
1-CA->poor,0.8275261324041812,950
1-Vi->poor,0.8261603375527427,979
AQ->poor,0.8249525616698292,1739
Vi->poor,0.824486301369863,1926
GE->poor,0.8227969348659003,1718
0-Vi->poor,0.8227628149435273,947
1-CP->poor,0.8223308883455582,1009
TA->poor,0.8197368421052632,1869
SG->poor,0.8192771084337349,1632
0->poor,0.8187117243414742,11096
Le->poor,0.8164948453608247,2376
1-SG->poor,0.8157129000969933,841
CP->poor,0.8129280821917808,1899
LI->poor,0.8121212121212121,2144
1-LI->poor,0.8039502560351134,1099
1-6081949->poor,0.7937080969571945,1539
```

图 3-0-10 sup=0.005 mincon=0.75 的关联规则

由 3.2.1 中的分析，除去星座因素，可初步得出性别、城市可能与用户的申购赎回行为存在关联，并且很容易得出结论，用户的贫富等级与其申购赎回之间也存在着关系。

3.2.3 基于用户信息和申购赎回数据的数据分析

基于 3.2.2 中的分析，项目将在此对数据进行分类划分，并实现可视化，进一步分析不同类别的用户对其申购赎回行为的影响关系。

注：以下所有的与金额有关的数据均为同组同类型的数据取均值，并分析其随时间演变而变化的结果。例如：城市的 tBalance 为在该所城市中所有用户的 tBalance 的均值。

1. 各城市的 tBalance-Time 图

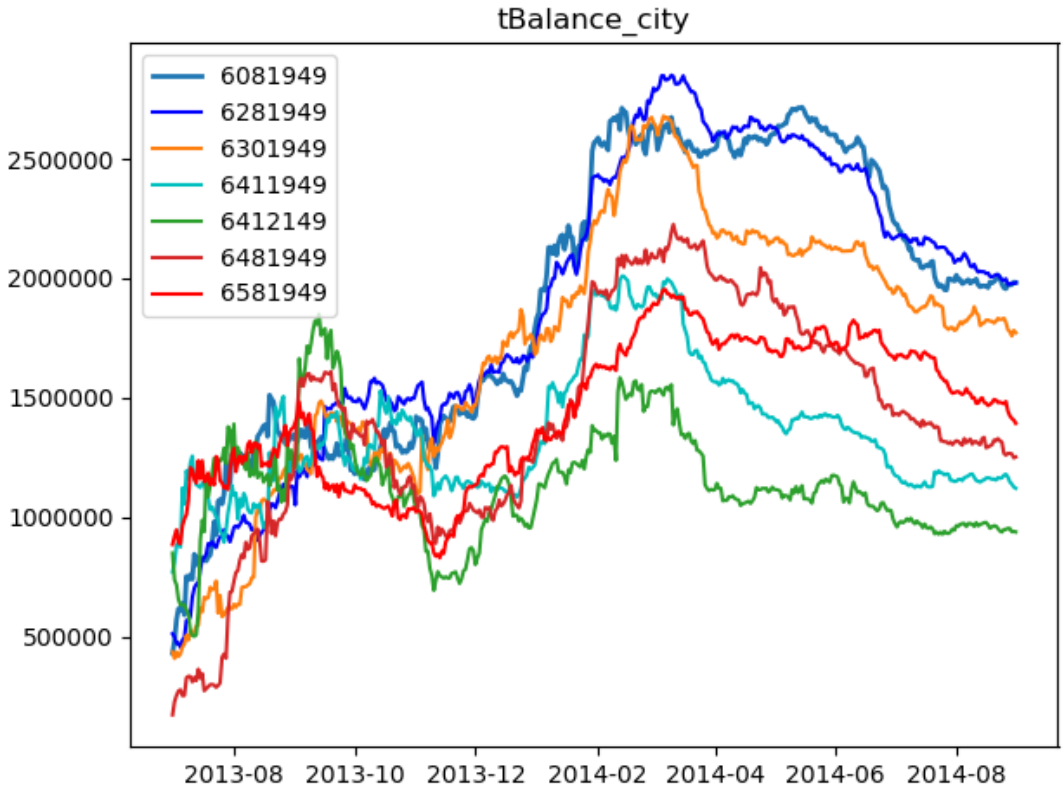


图 0-11 各城市的 tBalance-time 对比图

原始图像数据分布较为密集，故将图像横向拉伸，以更好的进行分析：

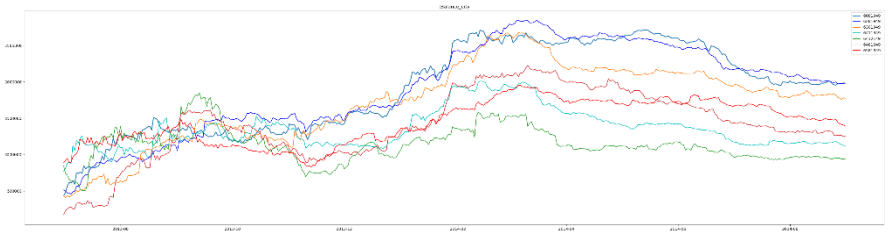


图 0-12 拉伸后的图 3-11

由拉伸后的图像克制，201403 前的数据曲线波动较大。

2. 各城市的 purchase-time/redeem-time 图

根据上述图像，将七个城市分为两类，（6081949、6281949、6301949 为一组，6411949、6412149、6481949、6581949 为一组），分别得到其与 purchase 和 redeem 的关系，数据可视化如图

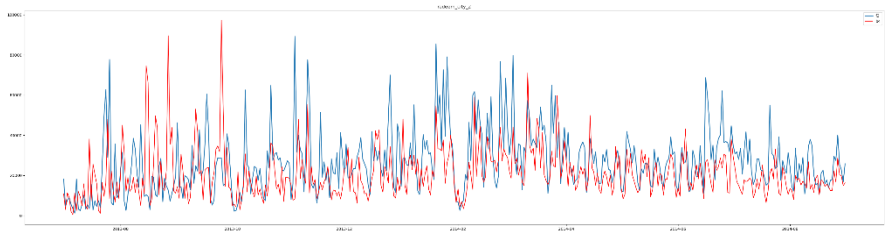


图 0-13 各城市的 purchase-time 对比图

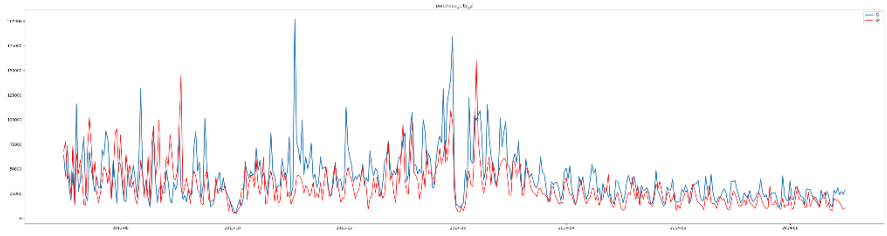


图 0-14 各城市的 redeem-time 对比图

3. 男性/女性的 purchase-Time/redeem-Time/tBalance-Time 图

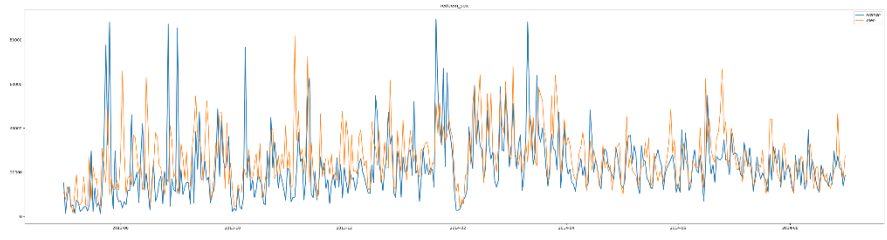


图 0-15 男/女的 purchase-time 对比图

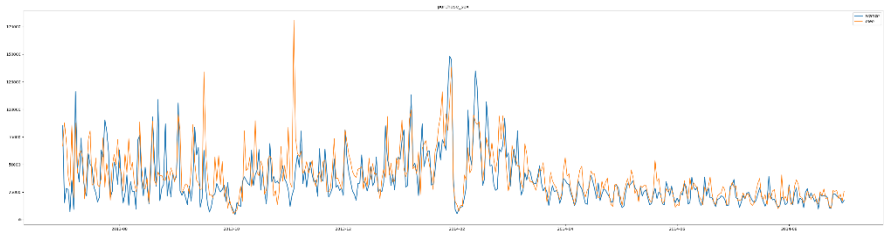


图 0-16 男/女的 redeem-time 对比图

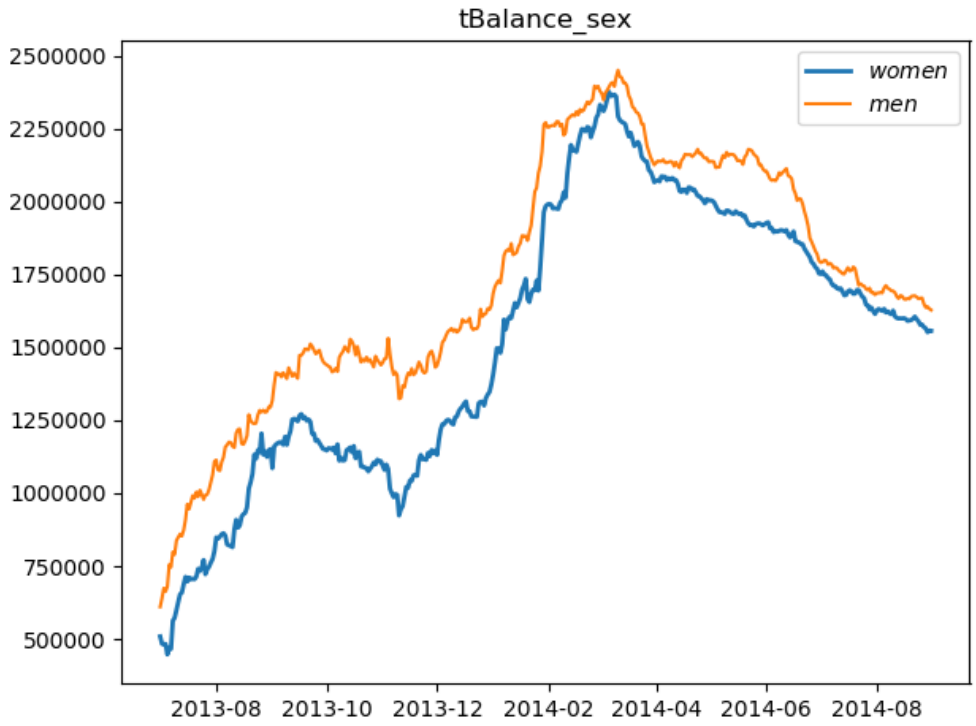


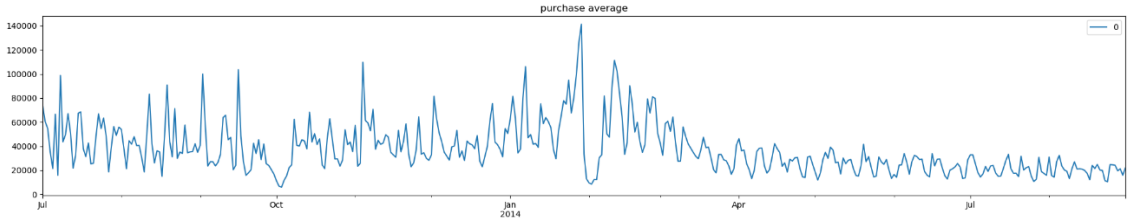
图 0-17 男/女的 tBalance-time 对比图

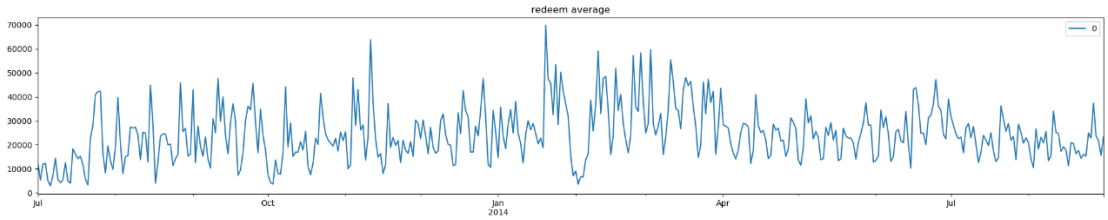
由图可知性别不同对于相关数据的整体趋势影响不大，但对具体数据值的大小略有影响

3.3 技术路线

3.3.1 用人均替代总额的求解

在该项目中，通过对数据的初步分析，可以得出用户人数与目标序列增长趋势的相似性，通过将每日的申购赎回总和除以每日的用户总量，可以得出每日人均申购赎回额，时序图如下，





可以看出，时序的趋势已被很大程度消除，更具平稳性。

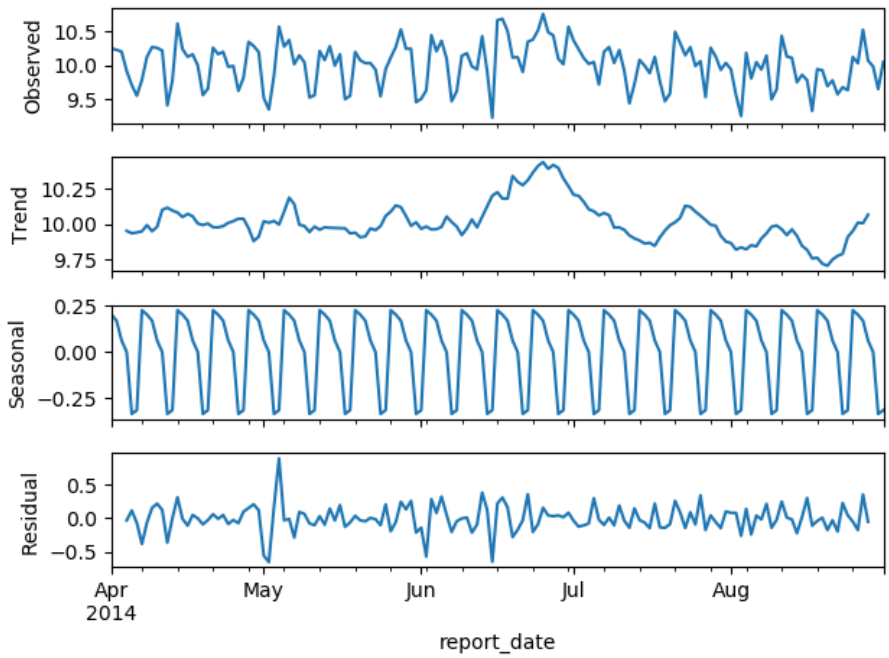
因此，在该项目中，核心目标由预测总额分解为两个更易求解的子目标——预测日用户总量和预测日平均申购赎回的。

3.3.2 STL 时间序列分解

由日均申购赎回图可知，日均时序仍具有不平稳性。本项目选择STL时间序列分解，以将时序转为可预测的平稳序列，并且根据该序列方差呈现随数值增大而增大的特点，得出乘法模型比加法模型更适合本项目的结论，公式如下，其中 St 是季节项，Tt 是趋势项，Rt 是剩余项。

$$y_t = S_t \times T_t \times R_t.$$

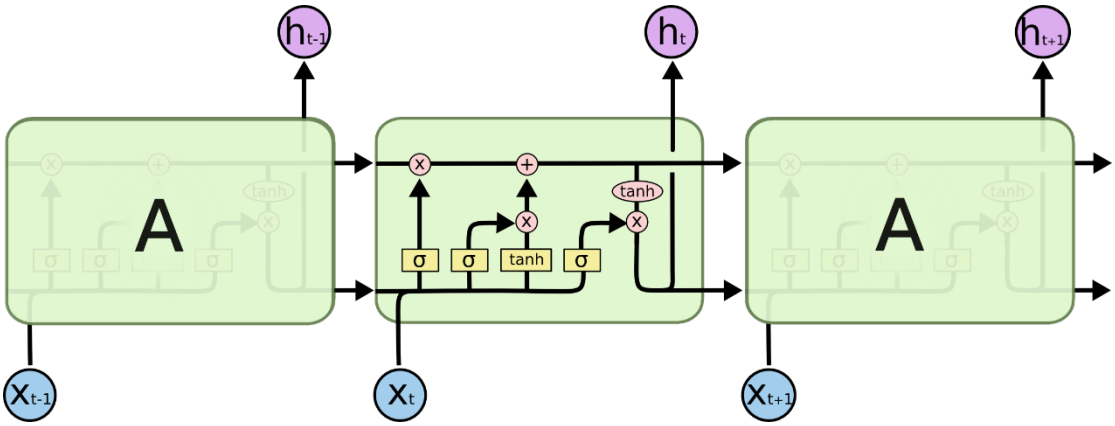
根据数据为按天序列的特点，我们以星期为周期进行 STL 时间序列分解，分解为趋势项、季节项和余项，如下。



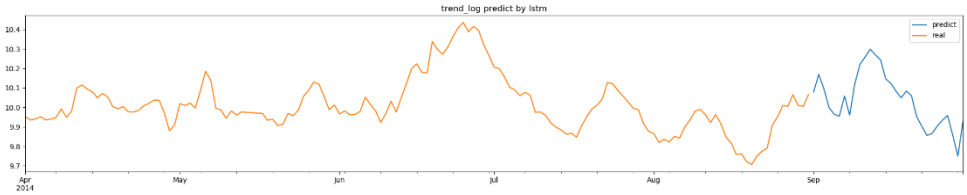
由图可知，2014 年 3 月前数据受春节与 2 月长假影响偏离过大，可能对预测结果带来额外不利的影响，因此本项目只选取时序中相对稳定的部分——2014 年 4 月到 8 月，进行分析预测。

3.3.3 Lstm 长短时间神经网络

由于 Lstm 可以有选择性的筛选出重要的信息，并通过将前面的信息存储下来的方式，计算后面的信息的特点，与趋势项具有长期性、受历史值影响的特点相吻合，本项目采用 Lstm 模型进行趋势项的预测与拟合，神经网络结构如下。



预测结果如下，



预测结果体现出，Lstm 已经可以初步学习到趋势项的变化规律和趋势。

3.3.5 多元线性回归

由于数据集比较短，并且将时间序列进行了人均处理，在该前提下，可以把余项看作是用户受特殊事件或时间点影响的群体行为，通过初步数据分析的异常值分析，和对余项中极值的比较，可以将余项看作是清明节、五一、中秋节、对正常工作日和周末带来的影响，由于节假日处于不同时间点，可能对该时间点及其前后时间段的用户行为造成不同的影响，本项目设计了如下时间特征：

1.是否是常规周（没有节假日）的第 1 天，是否是常规周末第 1 天，是否是假期的第 1 天，是否是假期前/后的第一天？

2.上班前一天是否休假，是否是工作日，是否是假期，是否是月初、月中或月末（1-10 日、11-20 日和 21-30 日，三个 01 特征），是否是每月第一天？

3.上一个波峰或波谷是几天前？

4. 上班最后天后要放几天假（2 天、3 天和 7 天，3 个 01 特征）？

5. 上班第一天前放了几天假（2 天和 3 天 2 个 01 特征）

6. 是否是周末（2 天假），是否是节假日（3 天假）？

7. 周末是否补班？

由于特征全部为 01 值，所以多项式回归与线性回归结果本质相同，并且也不必要使用岭回归或者套索回归，所以本项目采用多元线性回归分析的方式去拟合预测余项。公式如下，

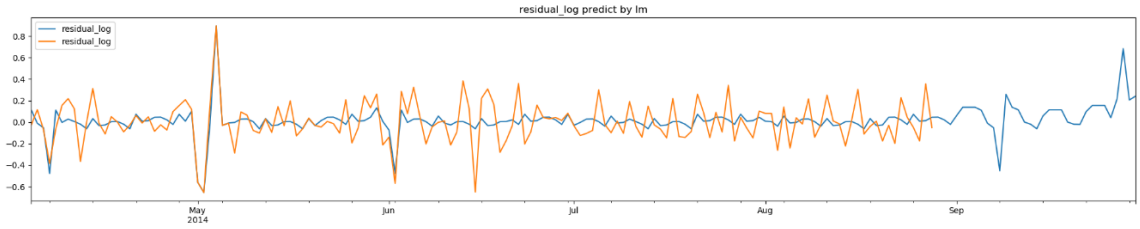
$$y_t = \beta_0 + \beta_1 x_{1,t} + \beta_2 x_{2,t} + \cdots + \beta_k x_{k,t} + \varepsilon_t,$$

然而这些时间特征都是类别变量，无法被机器理解，所以本项目首先对这些特征进行了 onehot 编码，全部处理为互斥的独立变量。

拟合过程中使用最小二乘法，最小化每个数据点与预测直线的垂直误差的平方和来计算得到最佳拟合曲线，并使用标准误差度量拟合优度指标，公式如下，

$$\sum_{t=1}^{T} \varepsilon_t^2 = \sum_{t=1}^{T} (y_t - \beta_0 - \beta_1 x_{1,t} - \beta_2 x_{2,t} - \cdots - \beta_k x_{k,t})^2.$$

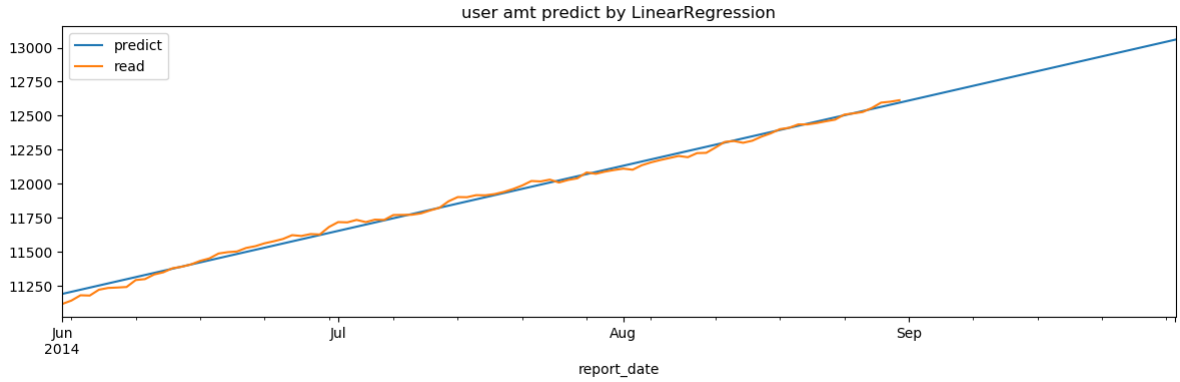
经过完全共线性、关联度的分析，与特征的筛选，最后预测结果如下，



通过训练的结果得知，模型可以学习到一定的节假日信息，而且由于可以通过标注 9 月的时间点与特殊节假日类别，起到辅助预测 9 月波动的作用，这在没有任何给定的 9 月依赖值，如余额宝银行利率下的情况下，这样可事先确定的因变量就格外重要。

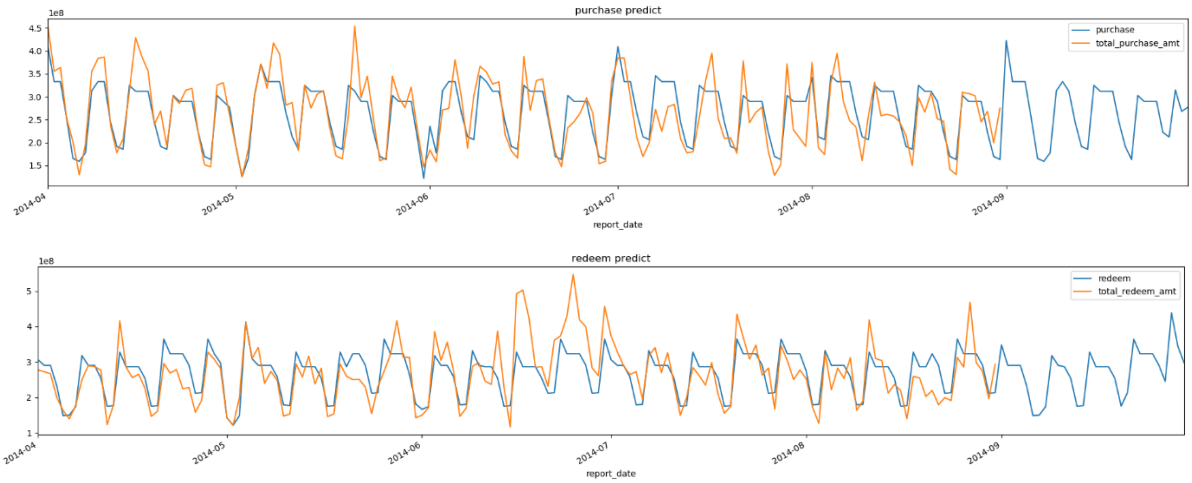
3.3.6 一元线性回归

由初步数据分析可知，用户数量再后期具有非常好的线性度，R^2 检验达到 0.95 之高，适合用一元线性回归进行拟合预测，对用户数量的预测结果如下，



3.3.7 模型组合

经过 STL 分解后三部分各部分分别的建模与预测，我们将各部分相乘，再经过人数的相乘后，最终日总和结果如下，



3.3.7 不足与优化方案

本项目的优化方案之一，可以从用户群体的划分着手。将申购和赎回经过人均后处理，分析时序图发现，时序图的方差波动前期大于后期，原因可能是由于用户数量上升而带来的整体申购赎回行为的稳定，但也可能是由于用户群体组成部分的改变而带来的影响，同样的，时序趋势的下降趋势也可能跟此有所关联。

除此之外，本项目还可以通过引入外部因素，训练多维的 Lstm 神经网络以预测趋势。目前趋势通过 Lstm 单维神经网络训练预测，但单维具有只能学习到内生变量——历史时间点值对未来时间点值的关联内生变量的缺点，无法分析除外生变量对数值可能会造成的影响，而经过初步的数据分析，我们可知，余额宝与银行利率完全有可能是造成趋势波动变化的因变量。

最后，本项目还可以通过划分赎回的资金流向以提高赎回预测的精准度。由于赎回具有两种资金流向——消费与转出到银行卡，这两者间，可能消费受周末与节假日的影响更多，而转出可能与其他因素关联更密切，本项目可以分别入手，分析两种流向不同的时序，已达到最佳的可解释性与精确度。

四、 市场分析

4.1 市场分析

4.1.1 外包平台分析

根据市场现状，企业需求方类型分为：关注价格的成本节约型、注重效率的高效型、关注服务方真实水平和人品的高质型。而个体/团队服务方分为关注实际收入的收益型、关注具体需求描述的成长型、要求高效率的效率型以及关注需求方综合实力的人际交往型。

4.1.2 技术需求定位

以余额宝为代表的可随存随取的基金旨在提供一种高流动性，低风险，但却能获得相对高收益的理财产品。其目的是吸引大量年轻时尚的消费者以及较为厌恶风险的散户投资者。

于是准确的预测资金的流入流出趋势（现金流）就对于相关客户十分重要。同时，由于此类基金可随存随取的特点，每日操作的频繁性，使用的普及性，使现金流的预测有一定难度的同时伴随着巨大的需求。

市场对于以余额宝为代表的可随存随取的基金的现金流、买入卖出进行较为准确地预测是需求巨大的。

则资金预测服务机构的市场定位为立足于个体/团队服务方的角度面向大型基金日现金流的预测，以及可尝试向小型电商、金融机构提供精准个性化预测服务。

4.2 市场中竞争者分析

现金流预测服务在其市场中的竞争者主要有两类：投行评估、公司内部财部门。

投行评估：投行作为第三方可以提供对委托公司的资金流预测服务，但一般收费较贵，分析周期长。投行体量庞大，有固定流程，结果客观但对委托公司并非十分保密。

公司内部财部门：公司内部财管部门对于公司现金流的预测一般比较私密，得到的信息更加全面，但可能不够客观公平，对企业各部门压力过大，耗费精力过多。

4.3 市场竞争力分析

通过对于市场中竞争者的分析，可知现金流趋势预测面临的竞争主要来自三个方面：潜在竞争者的进入风险、同业竞争者的竞争、替代品的威胁。现金流趋势预测在三方面威胁面前的竞争优势，构成了其在市场上的竞争能力。

4.3.1 潜在竞争者的进入风险

潜在竞争者是指那些目前尚未参与本行业竞争，但一旦他们选择参与将有能力参与竞争的公司或机构。例如可能计划推出资金流预测服务的小型投行。

而潜在竞争者进入风险与进入壁垒强弱正相关，进入壁垒越高则潜在竞争者进入市场所要承担的成本就越大。进入壁垒包括规模经济、品牌忠诚、绝对成本优势、顾客转换成本以及政府管制。在余额宝所处的金融产品市场中，对于资金预测服务的边际成本比较高，对于技术和客户都存在一定程度的依赖，因而对进入者的要求较高；顾客的转换成本也比较高。但是政府对该市场监管严格；传统小型预测分析服务机构所打造的品牌赢得了许多忠实客户，加上技术的熟练考虑，在这方面小型资金预测服务机构有相当大的优势。

总体来看金融产品市场有一定程度的进入壁垒，潜在竞争者的威胁不大，小型资金预测服务机构在这方面的竞争占优。

4.3.2 同业竞争者间的竞争

同业竞争是指市场内不同公司为了获得竞争对手的市场份额而进行的竞争性对抗。这种对抗主要通过价格、产品设计、广告宣传投入以及服务来实现。激烈的市场竞争会导致产品或服务只能降低价格，或同时导致企业必须支付更多的资金用于竞争。最终导致利润下降。

在余额宝所处的个人金融理财产品市场中，产品种类繁多、产业结构分散，不存在完全垄断或是寡头垄断。目前整个产业还处于成长期，随着社会财富的积累，整个市场还在逐步扩大，新增客户以及原有客户的需求不断增加。于是针对此类产业的资金流的预测服务，需求量巨大且必要，服务的固定成本低但边际成本高。因此说来市场中竞争者间的竞争并不局限于互相争夺市场，还在于谁能更好的提供服务，同时发现新的利润源。这对小型资金预测服务机构是十分有利的

4.3.3 **替代品的威胁**

替代品是指那些可以满足相同顾客的不同产品。因为相似替代品限制了资金流预测服务的定价和收益，相似替代品的存在给小型资金预测服务机构带来威胁。只有服务水准差异是对抗代替品威胁的手段。而小型资金预测服务机构最突出的差异在于其服务以及技术水准。因此小型资金预测服务机构通过提供更好的服务以维持客户忠诚，才是能创造更高竞争力的最佳途径。

五、 商业模式与发展战略

5.1 目标客户

5.1.1 成为证券投资分析全方位解决方案的提供商

向个人投资者提供互联网版、证券投资图文分析软件版；向券商机构提供局域网版；向银行提供外汇分析软件；向期货交易所、营业部提供期货分析软件；向海外投资人提供海外版证券分析软件；提供股票、期货、外汇“大集成”移动交易系统。

5.1.2 组建或收购投资咨询公司

在条件成熟后，向投资咨询机构、基金管理公司及个人投资者提供项目的技术分析资讯； 向电视台、专业期刊提供项目相关专栏资讯；向手机资讯运营商提供专业资讯。

5.2 盈利模式

5.2.1 提供证券投资顾问的收入模式

为客户提供股票、国债等的走势预测或电商、金融公司的资金预测，分析投资决策，收取咨询费用。

5.2.2 提供周期性的财务管理服务

主要为企业提供内部周期性的财务预测，为企业财务管理提供科学的指导，避免财务紧缩或财务赤字的出现。

5.2.3 软件界面广告

在软件流量有一定规模之后，通过广告位招商获得利润。

5.2.4 提供证券投资分析软件

向有需要的个人、组织或机构有偿提供证券、期货投资分析软件以获得报酬。

5.3 财务分析

5.3.1 项目开发所需硬件及成本

类别	名称	描述	型号	数量	单价	总价
服务器	IBM 机架式服务器	标配四个 Intel 六核 Xeon E7530 处理器，可扩至八路处理器，标配 4 块内存	BM System* 3850 X5（7145N04）	1	125000	125000
防火墙	思科防火墙	最高 300 网络吞吐量，3 个快速以太网端口，130000 的并发连数	CISCO ASA510-K8	1	15000	15000
阵列镜像软件	磁盘阵列的同步镜像软件	实现数据实时同时镜像备份	BM/Netfinity XP400	2	30000	60000
VPN	适用于小型网络或分支机	适用于小型网络或分支机构，局域网	深信服 S5100	1	7000	7000
光纤交换机	IBM 光纤交换机	光纤传输速度快，抗干扰能力强	691810X	2	17000	34000

5.3.2 人力成本

岗位	雇佣成本/月	人次	耗时/月	小计
需求分析师	8000	2	2	32000
项目经理	10000	1	2	20000
技术支持顾问	20000	1	2	40000
程序开发人员	8000	5	2	80000
测试人员	8000	2	1	16000
辅助人员	5000	2	2	20000
合计				208000

5.3.3 其他成本预算

类型	金额	备注
材料费	10000	如 A4 打印纸等耗材购置品
通讯费	10000	租用公共线路以及员工通讯费、交通费
专有技术购置费	20000	包括相关专利的申请和购买等
后期维护费用	20000	含人工费、材料费、固定资产折 旧费、审计费、系统维护费等
场地租用	20000	团队活动场地租用

5.3.4 预算总计

名称	金额
硬件设备	241000
人工成本	208000
其他成本	80000

项目总预算	529000

5.3.5 项目报价

取项目利润率 30%,风险基金率 15%，假设税率为 0,则项目报价为：

项目报价=项目总成本*（1+15%+30%）= 767050.00 元。

5.4 发展计划

5.4.1 建立企业内部的大数据资金预测应用平台

5.4.2 成为证券投资分析的全方位解决方案提供商、分公司分销商遍布各地、互联网上最热门的证券咨询网站、投资咨询大师的培训中心及权威资格的认定机构。

5.4.3 设立海外公司或办事处，向国外出口英文版本产品,同时成为国外投资者分析中国股市、期市和汇市的首选软件。

5.5 SWOT 分析

5.5.1 优势（Strengths）

1.产品技术领先，专有技术壁垒，一定程度上降低了被抄袭的可能性，或者说抄袭带来的收益远低于付出的成本。

2.销售渠道广泛、合作者众多，潜在市场大。

3.售后服务好，对于本公司售出的产品，我们承诺终身免费提供技术支持和故障分析。如有升级需要，我们仅收取软件升级部分的费用。

4.完善的回访机制。软件产品在交付对方使用一个月内，我们不断跟踪客户进行使用培训、使用体验调查、客户意见反馈等活动，和客户建立良好的伙伴关系。

5.5.2 劣势（Weaknesses）

1. 根据国际经验，合理的软件人才结构应该是软件蓝领、软件工程师、软件架构师与系统分析师并存的金字塔形状。人才基数由大到小形成梯次，他们之间的比例应该是 7:4:1。而我们公司目前没有达到这样一个合理的人才结构。通常程序员和工程师没有严格区分职责。

2.缺乏明析的主流客户群定位，尚没有自己的数据源。由于是特定领域的专业软件，所以新用户上手需要培训。另外，作为中小企业，我们的品牌形象较弱，认可度不高。

3. 缺乏关键核心技术。我们的软件产品处于国际软件产业价值链体系的中端，高端软件市场几乎被国外产品垄断。另外我们的产品只面向特定群体，市场的稳定性相对不高。作为对比，我们开发的应用，无不运行在基础软件平台上，如微软的 Windows，甲骨文的 oracle 数据库系统等。

5.5.3 机遇（Opportunities）

1.市场扩展空间大，股票市场四年周期走牛。而且随着经济发展，信息化正在渗透到生活的方方面面、互联网用户快速增加，可借助互联网快速传播，所以潜在用户群体基数很大。这样也就使得潜在的市场非常广阔。

2.更重要的是，政府对软件行业的重视程度增强。在“全国软件创新工作会议”上，科技部提出，我国软件产业应尽快转变“以模仿、跟踪和应用产品开发为主”的路线，实施“前沿突破、市场牵引、组织创新”的新战略。有了国家的保障，泛软件企业在制定发展战略的时候，不再是“无米之炊”，并且对融资和外部竞争的压力上有所缓解。

5.5.4 威胁（Threats）

1.知识产权保护不到位，软件行业是一个特殊的行业，软件属于无形资产，但软件的盗版成本很低，因此软件行业的在市场无序、假冒伪劣商品猖獗时，软件企业的产品更易于假冒。一旦发生侵权案件，就可能使软件企业产生“灭顶之灾”。侵权行为很猖獗。由于软件行业的侵权行为泛滥，严重挫伤开发企业的积极性。虽然像我们这类公司开发的行业软件遭到盗版的可能性不及系统软件，但是不乏一些小企业抄袭别人的软件技术，盗取他人的设计方案。

2.股票市场的波动。众所周知，股票市场波动是不稳定且无规律可寻的，所以一旦在某段时间里股票市场持续低迷，将会严重威胁到企业的发展甚至生存。

六、 风险与控制

6.1 市场风险

大数据应用的核心是通过对海量数据的收集和整理，通过数据模型和数据挖掘来满足客户对市场的预测、对价格的预测、对投资的决策等需要。所以随着市场的变化，人们对预测领域的需求会发生变化，这就需要项目团队后续打算构建的智能挖掘系统要不断开发出适应市场变化的数据挖掘模型。通过各种渠道不断收集客户新的预测趋势，同时建立专项的数据模型研发团队，来动态防范这种市场风险。

6.2 技术风险

在当今只是和信息爆炸的时代，科技发展日新月异，所以技术存在被淘汰或者缺少竞争力的风险，同时存在人才缺失的风险。为了防范这个风险，项目团队在设计技术架构时充分考虑架构的扩展性和灵活性，与时俱进不断融合当代新技术。同时，由于技术人员将直接影响项目核心竞争力的高低，所以项目团队需要建立完善的技术培训体系，不断加强对现有成员的技术培训。

6.3 资金管理风险

资金因素历来是影响项目建设的重要因素，本项目未来可能涉及的投资金额比较大，资金筹措以风险融资为主要方式，未来各项资金是否落实到位、科学管理、合理运用，将直接影响本项目的顺利实施。因此，必须认真落实各项工作，保证资金到位并进行科学管理，否则难以实现预定目标。

6.4 人才风险

随着行业竞争格局的不断演化，对人才的争夺势必将日趋激烈，如果项目团队未来不能在发展前景、薪酬、福利、工作环境等方面持续提供具有竞争力的待遇和激励机制，可能会造成人才队伍的不稳定。因此，项目团队必须营造吸引人才和稳定人才的机制，给成员提供良好的发展空间。同时增强成员的主人公意识，将个人利益与团队发展结合起来。

6.5 财务风险

财务风险主要是由财务人员自身素质制约及财务管理制度不完善造成的。所以项目团队计划引进高水平的财务管理人员，聘请会计师事务所检查团队财务内部控制制度，将潜在的风险控制在尽可能小的范围内。

七、 风险与控制

7.1 经济可行性

7.1.1 定量收益

合同/收款金额；

准确预测资金流系统为企业带来的预计投资损失降低

准确预测资金流系统为企业带来的预计投资价值增值

其他如从多余设备回收的收入等

7.1.2 非定量收益

预测服务的改进；

由资金流预估值错误操作引起的风险减少；

资金流投资差错的减少；

资金利用灵活性的增加；

资金流信息掌握情况的改进；

从以上收益分析可知， 该软件为企业带来的收益是长远的。

7.2 技术可行性

项目成员皆为软件工程系在读本科生，具有良好的理论基础和代码基础，学习能力强，创新能力高 。同时指导老师为学院教授，能提供较好的指导和硬件条件，为项目的成功开展提供了保证。

项目目前预测趋势良好，预测分数较高，并且在现有基础上依然具有较大的提升和优化空间，有良好的发展前景。

7.3 社会可行性

项目开发过程遵守各项法律法规，并且在项目服务开始时提供的用户条例明确地列出用户与项目服务之间的责任与利益关系，在问题发生时候能够充分的依据协定进行处理。

当与第三方公司进行合作时，会通过协议，明确双方之间的利益关系，做到处理利益关系时，有法可依。

7.4 管理和操作的可行性

本项目服务由专业人员负责，维护资金流等企业营运信息的保密性、安全性。专业人员按照服务流程提供专业资金预测服务，无需用户企业花费过多时间精力，整套服务流程清晰、简介，易于操作、便于管理。

7.5 可行性分析结论

本项目具有良好的经济效益和社会效益，符合当前国家政策要求和市场需求快速扩展的需要。以上从经济、技术、社会可行性的分析中可以看出，该资金流预测技术服务是可行的。