二〇一八年十二月十七日星期一

下午2时20分

一周未进行实验，今天重新启动，并力争年底前有所突破！

为了理清思路，首先进行几项分析工作：CERT5.2中2000个用户OCEAN决定的CPB倾向，即依据已有公式计算其CPB-I与CPB-O，并在此基础上由高到底进行两个列排序，分别输出三类Insiders的排序位置。

统计结果

开始计算CPB分数..

首先是CPB-I分数

**Insiders\_1 is : 30个用户中，12个低于一半，18个排序在1000以上，CPBO中9个低于1000**

**Insiders\_2 is : 30个用户17个在排序1000之后，13个在1000之上，CPBO中12个低于1000**

**Insiders\_3 is : 10个用户中2个在1000之后，8个在1000以上，CPBO中一样**

KEW0198 KEW0198 [-16.94, -21.806] 1234

DAS1320 DAS1320 [-20.328000000000003, -29.0792] 1687

GFM1815 GFM1815 [-7.26, -25.094] 229

EPG1196 EPG1196 [-19.36, -21.504] 1590

KBC1390 KBC1390 [-5.808, -17.8912] 83

PBC0077 PBC0077 [-9.68, -14.392000000000001] 507

SAF1942 SAF1942 [-6.776000000000001, -23.906399999999998] 160

ALT1465 ALT1465 [-5.808, -6.9712000000000005] 79

SLL0193 SLL0193 [-9.68, -20.112] 574

IHC0561 IHC0561 [-18.876, -29.6764] 1493

JKB0287 JKB0287 [-7.744, -22.6416] 267

DNJ0740 DNJ0740 [-8.228000000000002, -23.3092] 338

MIB0203 MIB0203 [-10.164000000000001, -22.3396] 613

REF1924 REF1924 [-12.1, -21.89] 927

AYG1697 AYG1697 [-15.004000000000001, -26.415599999999998] 1107

ISW0738 ISW0738 [-10.164000000000001, -16.0996] 644

FZG0389 FZG0389 [-22.264, -30.709600000000002] 1915

ZKP0542 ZKP0542 [-11.132, -27.3148] 782

PTH0005 PTH0005 [-14.52, -14.308] 1088

ALW0764 ALW0764 [-23.232, -14.884800000000002] 1955

ELT1370 ELT1370 [-15.488, -18.7632] 1121

NIV1608 NIV1608 [-10.164000000000001, -20.2596] 592

JUP1472 JUP1472 [-18.876, -17.7164] 1496

WHB1247 WHB1247 [-8.712, -13.056799999999999] 417

TMC0934 TMC0934 [-8.712, -20.3368] 415

WSK1857 WSK1857 [-19.36, -15.264] 1559

ETW0002 ETW0002 [-10.164000000000001, -12.459599999999998] 614

WWW0701 WWW0701 [-12.1, -12.01] 904

VAH1292 VAH1292 [-12.1, -15.65] 883

**Insiders\_2 is : 30个用户17个在排序1000之后，13个在1000之上**

VCF1602 VCF1602 [-20.328000000000003, -17.639200000000002] 1664

CKP0630 CKP0630 [-24.2, -20.9] 1977

ZIE0741 ZIE0741 [-19.36, -26.704] 1575

SIS0042 SIS0042 [-20.328000000000003, -28.559199999999997] 1676

TNB1616 TNB1616 [-20.812, -26.1068] 1771

TRC1838 TRC1838 [-14.52, -20.028] 1064

MDS0680 MDS0680 [-17.424, -19.3536] 1283

WDT1634 WDT1634 [-24.2, -14.14] 1997

OSS1463 OSS1463 [-15.488, -15.6432] 1114

CIF1430 CIF1430 [-11.616, -22.782400000000003] 835

MCP0611 MCP0611 [-21.78, -16.522] 1894

CHP1711 CHP1711 [-6.776000000000001, -15.0664] 155

GWG0497 GWG0497 [-11.616, -21.7424] 853

KSS1005 KSS1005 [-8.228000000000002, -21.2292] 326

NAH1366 NAH1366 [-6.292, -25.318800000000003] 132

RRS0056 RRS0056 [-8.228000000000002, -25.389200000000002] 308

ICB1354 ICB1354 [-20.812, -22.4668] 1740

BYO1846 BYO1846 [-22.748, -21.4972] 1942

HXP0976 HXP0976 [-22.264, -23.9496] 1932

HMS1658 HMS1658 [-6.292, -15.438799999999999] 119

HIS1394 HIS1394 [-8.228000000000002, -24.8692] 331

LVF1626 LVF1626 [-16.94, -15.565999999999999] 1270

MGB1235 MGB1235 [-8.228000000000002, -8.2292] 344

DCC1119 DCC1119 [-17.424, -25.593600000000002] 1280

SNK1280 SNK1280 [-20.328000000000003, -13.479199999999999] 1699

ITA0159 ITA0159 [-11.132, -17.4348] 758

JAL0811 JAL0811 [-18.876, -26.0364] 1512

OKM1092 OKM1092 [-7.26, -19.894] 234

HSN0675 HSN0675 [-9.196000000000002, -10.6044] 469

TMT0851 TMT0851 [-12.1, -23.970000000000002] 912

**Insiders\_3 is : 10个用户中2个在1000之后，8个在1000以上**

MPF0690 MPF0690 [-11.132, -9.1148] 764

CRD0272 CRD0272 [-12.1, -14.09] 879

VRP0267 VRP0267 [-11.616, -20.7024] 834

ELM1123 ELM1123 [-19.36, -26.704] 1599

GKW0043 GKW0043 [-8.712, -10.456800000000001] 399

ACA1126 ACA1126 [-4.84, -7.196000000000001] 26

KCM0466 KCM0466 [-13.068000000000001, -20.1052] 984

ZEH0685 ZEH0685 [-4.84, -20.716] 1

LAH0463 LAH0463 [-19.36, -15.783999999999999] 1602

CWW1120 CWW1120 [-11.616, -24.3424] 847

下午3时33分

上述实验分析发现，从CPB-O的角度对于场景三报复的用户有着较好的区分度，其他也多少有区分度，可见确实满意度作为一种心理变量驱动着数据行为，然而，依旧不准确。

从场景二入手

ZIE用户而言，分析其离职关联用户的离职时间

<<ZIE0741\_start>>:2010-08-27

2010-02:

**KSW0708,-1.0,0,[],0,0,1.0,[2010-02-04],3073968.0,2,0,1,**

**离职日期：2010-02-22，周一上班最后一天；**

2010-03:

**CDO0684,离职日期：2010-03-17 周三最后一天上班**

0.391304347826,16.0,[2010-01-04; 2010-01-05; 2010-01-08; 2010-01-12; 2010-01-13; 2010-01-15; 2010-01-18; 2010-01-20; 2010-01-21; 2010-01-22],27518.6875,0.0,7.0,[2010-01-05; 2010-01-07; 2010-01-08; 2010-01-11; 2010-01-13; 2010-01-15; 2010-01-19],1363520.0,10.0,10,7,

**NWP1609, 离职日期,2010-03-26，周五最后一天上班**

-1.0,0,[],0,0,1.0,[2010-02-09],27703.0,0.0,0,1,

2010-04:

2010-05:

**OCW1127,离职日期2010-05-10， 周一最后一天上班**

-1.0,0,[],0,0,1.0,[2010-01-08],34511.0,0.0,0,1,

2010-06:

**JXH1061, 离职日期：2010-06-23， 周三最后一天上班**

1.0,1.0,[2010-03-03],41853.0,0.0,0,[],0,0,1,0,

**TAG1610, 离职日期2010-06-02， 周三最后一天上班**

1.0,1.0,[2010-03-15],24317.0,0.0,0,[],0,0,1,0,

**BMR0865, 离职日期2010-06-04， 周五最后一天上班**

1.0,1.0,[2010-05-06],40914.0,0.0,0,[],0,0,1,0,

2010-07:

**BNS0484, 离职日期2010-07-21，周三最后一天上班**

1.0,11.0,[2010-01-06; 2010-01-08; 2010-01-12; 2010-01-13; 2010-01-14; 2010-01-18; 2010-01-19; 2010-01-22; 2010-06-22],30665.6363636,0.0,0,[],0,0,9,0,

**DAS1320, 离职日期2010-07-30， 周五最后一天上班**

1.0,1.0,[2010-07-30],25796.0,0.0,0,[],0,0,1,0,

**AWW0718, 离职日期2010-07-16， 周五最后一天上班**

-1.0,0,[],0,0,1.0,[2010-04-08],33501.0,0.0,0,1,

2010-08:

**ZJN1492, 离职日期：2010-08-11，周三最后一天上班**

1.0,1.0,[2010-01-20],18568.0,0.0,0,[],0,0,1,0,

CKP0630, 离职日期：2010-08-26， 周四最后一天上班

-1.0,0,[],0,0,1.0,[2010-03-15],2260635.0,2,0,1,

<<ZIE0741\_end>>

初步分析结果：

对于ZIE0741用户而言，其离职日期为2010-08-27，为当月的一个周五（最后一天登录）

VCF1602,2010-08-20, 周五离职

CKP0630,2010-08-26, 周四离职

SIS0042,2010-09-02, 周四离职

TNB1616,2010-09-10, 周五离职

TRC1838,2010-09-15, 周三离职

MDS0680,2010-09-17, 周五离职

WDT1634,2010-09-20, 周一离职

OSS1463,2010-09-21, 周二离职

CIF1430,2010-09-23, 周四离职

MCP0611,2010-10-06, 周三离职

CHP1711,2010-10-13, 周三离职

GWG0497,2010-10-15, 周五离职

KSS1005,2010-10-16, 周六离职

NAH1366,2010-11-17, 周三离职

RRS0056,2010-12-10, 周五离职

ICB1354,2010-12-15, 周三离职

BYO1846,2010-12-15, 周三离职

HXP0976,2010-12-20, 周一离职

HMS1658,2010-12-30, 周四离职

HIS1394,2010-12-30, 周五离职

LVF1626,2011-01-14, 周五离职

MGB1235,2011-01-21, 周五离职

DCC1119,2011-01-26, 周三离职

SNK1280,2011-02-11, 周五离职

ITA0159,2011-02-17, 周四离职

JAL0811,2011-02-25, 周五离职

OKM1092,2011-04-29, 周五离职

HSN0675,2011-04-29, 周五离职

TMT0851,2011-05-11, 周三离职

2018年12月18日星期二

上午9时56分

这里希望提出的假设是：可以根据离职的礼拜几来区分是解雇还是离职么？

**分析下场景三：全部是周五离职**

MPF0690,2010-06-18, 周五最后一天

CRD0272,2010-07-16, 周五离职

VRP0267,2010-08-08, 周日离职

ELM1123,2010-10-15, 周五离职

GKW0043,2010-11-05, 周五离职

ACA1126,2010-12-03, 周五离职

KCM0466,2011-02-11, 周五离职

ZEH0685,2011-04-01, 周五离职

LAH0463,2011-05-06, 周五离职

CWW1120,2011-05-13, 周五离职

继续分析场景一

KEW0198,2010-07-29, 周四离职

DAS1320,2010-07-30, 周五离职

GFM1815,2010-07-30, 周五离职

EPG1196,2010-08-05, 周四离职

KBC1390,2010-08-10, 周二离职

PBC0077,2010-08-13, 周五离职

SAF1942,2010-08-13, 周五离职

ALT1465,2010-09-08, 周三离职

SLL0193,2010-09-23, 周四离职

IHC0561,2010-09-28, 周二离职

JKB0287,2010-10-04, 周一离职

DNJ0740,2010-10-26, 周二离职

MIB0203,2010-11-04, 周四离职

REF1924,2010-11-10, 周三离职

AYG1697,2010-11-16, 周二离职

ISW0738,2010-11-22, 周一离职

FZG0389,2010-12-30, 周四离职

ZKP0542,2011-02-04, 周五离职

PTH0005,2011-02-18, 周五离职

ALW0764,2011-02-28, 周一离职

ELT1370,2011-03-01, 周二离职

NIV1608,2011-03-03, 周四离职

JUP1472,2011-03-09, 周三离职

WHB1247,2011-03-11, 周五离职

TMC0934,2011-03-14, 周一离职

WSK1857,2011-03-30, 周三离职

ETW0002,2011-04-04, 周一离职

WWW0701,2011-04-18, 周一离职

VAH1292,2011-05-26, 周四离职

感觉上似乎在离职日期上应该有规律，试着写个程序分析下CERT5.2中离职用户的礼拜几。

注意一年中31天的月份有一三五七八十腊，剩余月份2月闰年29天，平年28天（能被4整除但是不能被100整除的为闰年）

初步统计了CERT5.2中238个离职用户的离职礼拜天

1 25

2 33

3 50

4 37

5 87

6 2

7 4

上图中可以发现，出去周末两天，绝大部分离职人数在后半周离职，即周三-周五

因此，选择将前半周（周一周二）作为可能的解雇用户，后半周作为主动离职用户

重新统计CERT5.2中所有用户的离职情况：

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| User | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Leave\_1half | Leave\_BackHalf |

<<MTD0971\_start>>:2010-10-21 以该用户进行验证

MTD0971,1.0,1.0,3.0,4.0,6.0,0.0,0.0,2.0,13.0,

FDS1841,-1.0,0,[],0,0,1.0,[2010-01-08],2680277.0,1,0,1, 周三

ZAD1621,-1.0,0,[],0,0,1.0,[2010-01-22],609257.0,1,0,1, 周四

PKS1187,1.0,1.0,[2010-01-12],40344.0,0.0,0,[],0,0,1,0, 周五

EAL1813,1.0,1.0,[2010-01-04],687829.0,1,0,[],0,0,1,0, 周三

**GWH0961,BMR0865,1.0,1.0,[2010-06-02],59383.0,0.0,0,[],0,0,1,0, 周四**

**NWH0960 周四**

**BMR0865 周五**

KEW0198,0.0,1.0,[2010-07-12],40612.0,0.0,1.0,[2010-03-03],51171.0,0.0,1,1, 周四

ZJN1492,1.0,1.0,[2010-02-01],20682.0,0.0,0,[],0,0,1,0, 周三

ZHB1104,1.0,1.0,[2010-02-26],4252197.0,3,0,[],0,0,1,0, 周五

SCO1719,1.0,1.0,[2010-06-01],39971.0,0.0,0,[],0,0,1,0, 周五

ILG0879,1.0,1.0,[2010-03-03],37798.0,0.0,0,[],0,0,1,0, 周一

IVS1411,1.0,1.0,[2010-06-09],3016372.0,3,0,[],0,0,1,0, 周五

OSS1463,1.0,1.0,[2010-06-30],29841.0,0.0,0,[],0,0,1,0, 周二

MFM1400,1.0,1.0,[2010-01-07],69753.0,1,0,[],0,0,1,0, 周五

<<MTD0971\_end>>

验证通过！

直接分析2000个用户的Leave\_Users中前半段情况，然后使用随机森林计算区分场景二的特征分类度。

直接分析周一周二离职人数的情况，Insiders\_2不具有区分度；

分析周三-周五离职人数的情况，Insiders\_2同样不可区分；