二〇一九年一月二日星期三

上午10时28分

元旦节后开工。

早上提交了工程师转助研申请。

开始元旦假期中自己的实验假设，即基于静态的26JS与33维度的ATF的分析，主要分析对象为CERT5.2用户。

基于静态分析假设，更容易调整参数得到最好的结果。

先基于原有的26JS维度与新的33维度的ATF进行研究吧！

下午3时36分

分别按照传统的JS与ATF进行实验，首先，进行KMeans分析，可以得到：

26JS: [-1]用户可以分成2个群簇，分别包含

0 1564

1 197 （显然应该是我们特别关注的部分）

而使用33维度的ATF实验，有 针对ATF\_lst的

最佳自动K值为[2:10] 2 : 0.36697501148725215

0 197 个

1 1564 个

下面继续使用后续的ATF聚类结果进行实验，并且默认采用1564个作为训练集，剩余197个[-1]+[+1]作为测试集输出结果；

计算召回率与FPR

197/1564=0.12...

下面开始记录针对ATF特征的预测结果：

自动PCA（）

[**0.05**, 0.9, 0.47005241751310445, 0.5362318840579711, **0.12556278139069535**, 0.6651270207852193, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, **random\_state=15**, shrinking=True, tol=0.01,

verbose=False), **0.5172413793103449, 0.6, 0.4**]

[**0.01**, 1.0, 0.5362318840579711, 0.5362318840579711, **0.128064032016008**, 0.6766743648960739, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.01, **random\_state=15**, shrinking=True, tol=0.01,

verbose=False), **0.5172413793103449, 0.6, 0.4**]

[0.01, 1.0, 0.5362318840579711, 0.5362318840579711, **0.128064032016008**, 0.6766743648960739, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.01, **random\_state=0**, shrinking=True, tol=0.01,

verbose=False), **0.5172413793103449, 0.6, 0.4**]

[0.05, 1.0, 0.5362318840579711, 0.5362318840579711, 0.12556278139069535, 0.6651270207852193, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, **random\_state=5**, shrinking=True, tol=0.01,

verbose=False), **0.5172413793103449, 0.6, 0.4**]

[0.05, 1.0, 0.5362318840579711, 0.5362318840579711, 0.12556278139069535, 0.6651270207852193, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, **random\_state=10**, shrinking=True, tol=0.01,

verbose=False), **0.5172413793103449, 0.6, 0.4**]

[0.05, 1.0, 0.5362318840579711, 0.5362318840579711, 0.12556278139069535, 0.6651270207852193, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, **random\_state=100**, shrinking=True, tol=0.01,

verbose=False), **0.5172413793103449, 0.6, 0.4**]

下面开始调整PCA的维度

PCA=3

[0.05, 1.0, 0.6811594202898551, 0.6811594202898551, 0.13356678339169584, 0.7251732101616628, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, **random\_state=100**, shrinking=True, tol=0.01,

verbose=False), **0.6551724137931034, 0.6, 1.0**]

[0.05, 1.0, 0.6811594202898551, 0.6811594202898551, 0.13356678339169584, 0.7251732101616628, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, **random\_state=0**, shrinking=True, tol=0.01,

verbose=False), **0.6551724137931034, 0.6, 1.0**]

PCA=1

*[0.05, 1.0, 0.8695652173913043, 0.8695652173913043, 0.17588204318062137, 0.19709854927463732, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',*

*max\_iter=-1, nu=0.05, random\_state=10, shrinking=True, tol=0.01,*

*verbose=False), 0.7931034482758621, 0.9, 1.0]*

[0.05, 1.0, 0.8695652173913043, 0.8695652173913043, 0.16708354177088544, 0.9099307159353349, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, **random\_state=10**, shrinking=True, tol=0.01,

verbose=False), **0.7931034482758621, 0.9, 1.0**]

[0.15, 0.9, 0.6888400722100181, **0.782608695652174**, 0.15507753876938468, 0.8406466512702079, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.15, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **0.7586206896551724, 0.7666666666666667, 0.9**]

PCA=2

[0.05, 1.0, 0.7971014492753623, 0.7971014492753623, 0.14357178589294647, 0.789838337182448, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, **random\_state=10**, shrinking=True, tol=0.01,

verbose=False), **0.7931034482758621, 0.7666666666666667, 0.9**]

[0.05, 1.0, 0.8260869565217391, 0.8260869565217391, 0.1535767883941971, 0.8406466512702079, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto',

kernel='linear', max\_iter=-1, nu=0.05, random\_state=10,

shrinking=True, tol=0.01, verbose=False), 0.7931034482758621, 0.8, 1.0]

今天下午最大的结果就是，通过PCA=1的KMeans+OCSVM 可以得到了recall基本满意，FPR=0.167的结果

二〇一九年一月三日星期四

下午3时34分

今天下午继续昨天的实验，上午终于把此次工程师转助研的手续搞定。

昨天最好结果的实验中，分析DF函数的排序关系：一共433个用户，一半是：217

**Insiders\_1 in Risk\_Users\_Sort:...**

['KEW0198', 1, 1, '0.37430310077277085'] : 62

['DAS1320', -1, 1, '-22.23026921734859'] : 431

['GFM1815', 1, 1, '0.014400357126774566'] : 284

['EPG1196', 1, 1, '0.005637024245942257'] : 341

['KBC1390', -1, 1, '-0.0018274702185152591'] : 413

['PBC0077', -1, 1, '-0.0015464048043298817'] : 412

['SAF1942', -1, 1, '-22.14763131692173'] : 430

['ALT1465', 1, 1, '0.026804458461665348'] : 190

['SLL0193', 1, 1, '0.011109562879745738'] : 313

['IHC0561', 1, 1, '0.6022116960271156'] : 21

['JKB0287', 1, 1, '0.02678587283564937'] : 192

['DNJ0740', 1, 1, '0.004218641270885115'] : 353

['MIB0203', -1, 1, '-0.7209941744624366'] : 419

['REF1924', 1, 1, '0.015845453967635592'] : 273

['AYG1697', -1, 1, '-0.17277538219780197'] : 417

['ISW0738', 1, 1, '0.017838005871674056'] : 253

['FZG0389', 1, 1, '0.6268173201910905'] : 8

['ZKP0542', 1, 1, '0.0024564748586293206'] : 365

['PTH0005', 1, 1, '0.07336473924896225'] : 165

['ALW0764', 1, 1, '0.01843704214893549'] : 251

['ELT1370', 1, 1, '0.018709174248449756'] : 248

['NIV1608', 1, 1, '0.03132570395781187'] : 185

['JUP1472', 1, 1, '0.5830946344728041'] : 26

['WHB1247', 1, 1, '0.0266964563180494'] : 196

['TMC0934', 1, 1, '0.009238971102202953'] : 323

['WSK1857', 1, 1, '0.021208592547502292'] : 236

['ETW0002', 1, 1, '0.012960862896708392'] : 299

['WWW0701', 1, 1, '0.0006780833208104298'] : 390

['VAH1292', 1, 1, '0.6009325182006826'] : 23

**Insiders\_2 in Risk\_Users\_Sort:... 一半是217 如果严格按照217取一半，则只有16/30的召回率**

*['VCF1602', 1, 1, '0.0012927582339194998'] : 381*

['CKP0630', 1, 1, '0.12537746087586754'] : 139

['ZIE0741', 1, 1, '0.1621446298423379'] : 104

*['SIS0042', -1, 1, '-0.0019075955446155035'] : 415*

['TNB1616', 1, 1, '0.0147230069068911'] : 280

['TRC1838', 1, 1, '0.1733531135544517'] : 101

['MDS0680', 1, 1, '0.06374327823238346'] : 169

['WDT1634', 1, 1, '0.009124014033783823'] : 324

['OSS1463', 1, 1, '0.014395350212229374'] : 285

['CIF1430', 1, 1, '0.005506677019958772'] : 342

['MCP0611', 1, 1, '0.2563333269471464'] : 76

['CHP1711', 1, 1, '0.0552890363477907'] : 175

['GWG0497', 1, 1, '0.0618355714999268'] : 170

['KSS1005', 1, 1, '0.0021440004798805035'] : 366

['NAH1366', -1, 1, '-1.4854964916386066'] : 420 （全勤）

['RRS0056', 1, 1, '0.11198917020182719'] : 145

['ICB1354', 1, 1, '0.18827216216679687'] : 97

['BYO1846', 1, 1, '0.009371152081886436'] : 322

['HXP0976', 1, 1, '0.4857005567121533'] : 35

['HMS1658', 1, 1, '0.4246631586547167'] : 50

['HIS1394', 1, 1, '0.015579900707372474'] : 275

['LVF1626', 1, 1, '0.6099363019644386'] : 17

['MGB1235', 1, 1, '0.026757305975312562'] : 195

['DCC1119', 1, 1, '0.02350410810993253'] : 224

['SNK1280', -1, 1, '-0.00031667240468991054'] : 395

['ITA0159', 1, 1, '0.6138104688498807'] : 15

['JAL0811', 1, 1, '0.11014801702344101'] : 146

['OKM1092', 1, 1, '0.005071790815957655'] : 348

['HSN0675', 1, 1, '0.0005269671761425343'] : 391

['TMT0851', 1, 1, '0.0016731237254745679'] : 371

**Insiders\_3 in Risk\_Users\_Sort:...**

['MPF0690', 1, 1, '0.3594052219463215'] : 63

['CRD0272', 1, 1, '0.02171323050862739'] : 235

['VRP0267', 1, 1, '0.024757288365549357'] : 216

['ELM1123', 1, 1, '0.0036080763461114884'] : 359

['GKW0043', 1, 1, '0.026772242315434625'] : 194

['ACA1126', 1, 1, '0.026458459631438558'] : 203

['KCM0466', 1, 1, '0.13424506282532533'] : 135

['ZEH0685', 1, 1, '0.25570502196787004'] : 77

['LAH0463', 1, 1, '0.6248992267356286'] : 11

['CWW1120', 1, 1, '0.012644522112569945'] : 302

从上图中分析，如果以排序302作为分界线，即高于302之后的用户不再考虑，那么对于Insiders2而言Recall变化为(16+4)/30， Insiders3的Recall变化为9/10-->0.9，Insiders1的Recall为17/30,

而此时有302个用户则

Recall:

FPR: (302-46) / 2000 = 0.128 --- > 目标：FPR: 0.05-0.09

目前KMeans+OCSVM的最好结果PCA=1，

如果直接使用所有DF的中位数，即取前一半的RiskUsers，则有Insiders\_3的召回率迅速调到0.7，整体Risk比例为216 / 2000.0 = 0.108

Risk\_1结果中，433个测试用户中竟然识别Risk达到394个，其中误报太多了！

PCA=1

*[0.05, 1.0, 0.8695652173913043, 0.8695652173913043, 0.17588204318062137, 0.19709854927463732, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',*

*max\_iter=-1, nu=0.05, random\_state=10, shrinking=True, tol=0.01,*

*verbose=False), 0.7931034482758621, 0.9, 1.0]*

验证

Risk\_1\_ATF文件：

**HBW0057**,35.0,33.0,43.0,44.0,29.0,-16.99,-32.84,20.47,0.0,0.0,0.0,0.0,0.0,0.0,0.0,0.0,-12.26,-16.2,13.2053177459,0.293451505464,0.254693593253,0.00565985762785,-0.261707988981,5.02222222222,12.7555555556,428930.066327,136300.475979,3.46666666667,11.5777777778,2.97777777778,5.08888888889,5.91111111111,5.0,9.0,343.0,**0.0167852630153,**

**HBW0057,1,-1,0.016785263015268725,**

**验证通过！**

或许，在中低满意度群体上，出勤率可以作为一个有效的进一步筛选的表示器

初步对Risk\_1中判定为Risk的用户继续做自动KMeans，则有

Insider\_3: 9个一类，1个另一类

Insiders\_2: 除去3个未检测出，剩余又丢掉5个，总计检测22个一类

Insiders\_1中也有5个分在一类，剩余19个在另一类

356

38

如果把38个的一类去掉，变成356个Risk，占比17.8%， 且Recall降低明显

Insiders\_2的Recall从0.9-->0.73

Insiders\_3的Recall从1.0-->0.9

二〇一九年一月四日星期五

下午12时42分

下午2时46分

昨天对于Risk\_1的用户进行KMeans效果并不好，Recall降低且FPR降低有限。故我们今天尝试从late\_days, early\_days, cnt\_days来进行分析，主要依据scale,minmax两种依据，查看Insiders1/2/3的结果

先来看Insiders\_3的结果：

全周期的Late/Early占比约在0.2-0.4之间

insider\_3 MPF0690 [0.3898305084745763, 0.1864406779661017, 118.0] [0.38983051 0.36269334 0.20210526] [ 0.59505882 -0.34835138 -1.4560034 ]

insider\_3 CRD0272 [0.34057971014492755, 0.12318840579710146, 138.0] [0.34057971 0.2396452 0.24421053] [ 0.34745024 -0.66968901 -1.2729927 ]

insider\_3 VRP0267 [0.37089201877934275, 0.14553990610328638, 213.0] [0.37089202 0.28312681 0.40210526] [ 0.49984549 -0.5561377 -0.58670257]

insider\_3 ELM1123 [0.4527363184079602, 0.12437810945273632, 201.0] [0.45273632 0.2419596 0.37684211] [ 0.91131802 -0.66364501 -0.69650899]

insider\_3 GKW0043 [0.4027777777777778, 0.14814814814814814, 216.0] [0.40277778 0.28820077 0.40842105] [ 0.66015126 -0.54288717 -0.55925097]

insider\_3 ACA1126 [0.36324786324786323, 0.1282051282051282, 234.0] [0.36324786 0.24940451 0.44631579] [ 0.46141447 -0.64420278 -0.39454134]

insider\_3 KCM0466 [0.3665480427046263, 0.12099644128113879, 281.0] [0.36654804 0.23538106 0.54526316] [ 0.47800613 -0.68082474 0.0355338 ]

insider\_3 ZEH0685 [0.36908517350157727, 0.1608832807570978, 317.0] [0.36908517 0.31297513 0.62105263] [ 0.49076157 -0.47818946 0.36495306]

insider\_3 LAH0463 [0.37941176470588234, 0.1588235294117647, 340.0] [0.37941176 0.30896818 0.66947368] [ 0.54267854 -0.48865352 0.57541537]

insider\_3 CWW1120 [0.3710144927536232, 0.11884057971014493, 345.0] [0.37101449 0.23118714 0.68 ] [ 0.50046122 -0.69177707 0.62116804]

再来看看Insiders\_2

如果单独看缺勤率比例，则Insiders\_2中至少有12/27的用户会被筛选出去；

insider\_2 VCF1602 [0.36492890995260663, 0.45023696682464454, 211.0] [0.36492891 0.87587082 0.39789474] [ 0.46986593 0.99180082 -0.60500364]

insider\_2 CKP0630 [0.10843373493975904, 0.09036144578313253, 166.0] [0.10843373 0.17578511 0.30315789] [-0.81966454 -0.83645831 -1.01677772]

insider\_2 ZIE0741 [0.39520958083832336, 0.3592814371257485, 167.0] [0.39520958 0.69893001 0.30526316] [ 0.62210212 0.52972361 -1.00762718]

insider\_2 TNB1616 [0.36312849162011174, 0.4134078212290503, 179.0] [0.36312849 0.80422505 0.33052632] [ 0.46081432 0.8046994 -0.89782076]

insider\_2 TRC1838 [0.0, 0.0, 179.0] [0. 0. 0.33052632] [-1.36481555 -1.29551742 -0.89782076]

insider\_2 MDS0680 [0.0, 0.0, 181.0] [0. 0. 0.33473684] [-1.36481555 -1.29551742 -0.87951969]

insider\_2 WDT1634 [0.5, 0.47802197802197804, 182.0] [0.5 0.92992254 0.33684211] [ 1.14893636 1.13295574 -0.87036916]

insider\_2 OSS1463 [0.47549019607843135, 0.45588235294117646, 204.0] [0.4754902 0.8868531 0.38315789] [ 1.02571322 1.02048082 -0.66905739]

insider\_2 CIF1430 [0.43243243243243246, 0.3621621621621622, 185.0] [0.43243243 0.70453404 0.34315789] [ 0.80924015 0.54435842 -0.84291755]

insider\_2 MCP0611 [0.0, 0.00510204081632653, 196.0] [0. 0.00992528 0.36631579] [-1.36481555 -1.26959776 -0.74226167]

insider\_2 CHP1711 [0.0, 0.0, 199.0] [0. 0. 0.37263158] [-1.36481555 -1.29551742 -0.71481006]

insider\_2 GWG0497 [0.46766169154228854, 0.4527363184079602, 201.0] [0.46766169 0.88073295 0.37684211] [ 0.98635539 1.00449816 -0.69650899]

insider\_2 KSS1005 [0.011857707509881422, 0.05138339920948617, 253.0] [0.01185771 0.09995896 0.48631579] [-1.30520088 -1.03447669 -0.22068118]

insider\_2 RRS0056 [0.0, 0.0, 239.0] [0. 0. 0.45684211] [-1.36481555 -1.29551742 -0.34878867]

insider\_2 ICB1354 [0.0, 0.0, 242.0] [0. 0. 0.46315789] [-1.36481555 -1.29551742 -0.32133706]

insider\_2 BYO1846 [0.004132231404958678, 0.01652892561983471, 242.0] [0.00413223 0.03215463 0.46315789] [-1.34404075 -1.21154628 -0.32133706]

insider\_2 HXP0976 [0.46938775510204084, 0.3795918367346939, 245.0] [0.46938776 0.73844095 0.46947368] [ 0.99503318 0.6329056 -0.29388546]

insider\_2 HMS1658 [0.41832669322709165, 0.47410358565737054, 251.0] [0.41832669 0.92229987 0.48210526] [ 0.73832349 1.11304931 -0.23898225]

insider\_2 HIS1394 [0.350597609561753, 0.3545816733067729, 251.0] [0.35059761 0.6897873 0.48210526] [ 0.39781527 0.50584761 -0.23898225]

insider\_2 LVF1626 [0.4521072796934866, 0.45977011494252873, 261.0] [0.45210728 0.89441618 0.50315789] [ 0.90815552 1.04023164 -0.1474769 ]

insider\_2 MGB1235 [0.0037593984962406013, 0.0037593984962406013, 266.0] [0.0037594 0.00731337 0.51368421] [-1.34591516 -1.27641872 -0.10172422]

insider\_2 DCC1119 [0.3754646840148699, 0.14869888475836432, 269.0] [0.37546468 0.28927215 0.52 ] [ 0.52283458 -0.54008929 -0.07427262]

insider\_2 ITA0159 [0.0, 0.0, 285.0] [0. 0. 0.55368421] [-1.36481555 -1.29551742 0.07213594]

insider\_2 JAL0811 [0.010309278350515464, 0.030927835051546393, 291.0] [0.01030928 0.06016562 0.56631579] [-1.31298562 -1.13839616 0.12703915]

insider\_2 OKM1092 [0.4208955223880597, 0.4537313432835821, 335.0] [0.42089552 0.88266862 0.65894737] [0.75123829 1.00955314 0.52966269]

insider\_2 HSN0675 [0.0029850746268656717, 0.005970149253731343, 335.0] [0.00298507 0.01161406 0.65894737] [-1.34980808 -1.26518754 0.52966269]

insider\_2 TMT0851 [0.3699421965317919, 0.3786127167630058, 346.0] [0.3699422 0.73653621 0.68210526] [0.49507025 0.62793142 0.63031858]

如果我们直接定义成



则对于Insiders\_3而言，全部用户的LED指标都在0.3以上，准确的说是在0.3-0.4之间

对于Insiders\_2而言，

如果以0.30作为LED的下限，那么Risk\_1中有0.68的用户识别为Risk\_2，其中只有Insiders\_3作为Recall=1，此时

Risk\_2用户个数为: 248.0 0.629441624365 0.124 （Risk比例）

突然想到，如果使用不包含OCEAN的ATF数据呢？结果一样。

突然灵机一动，通过重新修改整体的使用过程，即KMeans选出的用户群簇用来训练OCSVM后直接测试Leave\_Users，依旧是PCA=1时，原始的Static\_ATF文件

[0.05, 1.0, 0.8695652173913043, 0.8695652173913043, 0.0832016850974197, 0.10905452726363181, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.7931034482758621, 0.9, 1.0

如此，FPR终于降到了0.08左右，而Recall则达到了最好的0.79/0.9/1.0

剩下的只需要使用DF的MinMax值进行预测即可，正常用户默认是0，Risk用户的DF值为+，越大威胁越高；

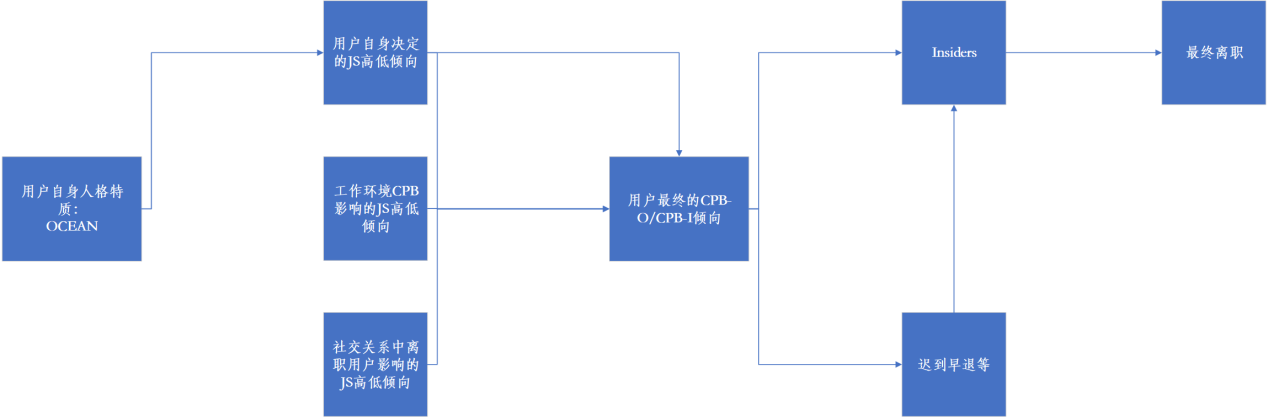
此时的238个LeaveUsers中有218个用户识别为满意度低用户，剩余20个用户为中等满意度；验证了我们的猜想：工作低满意度最终导致的离职，但是离职不一定原因是低满意度；

二〇一九年一月五日星期六

下午7时35分

2019年1月第1周，预感即将对于毕业论文中的主观建模部分有个交代。

目前的模型情况：



上图中，建模的重点是用户的反生产行为CPB作为最终Insider Risk风险，而CPB来源于自身的OCEAN特质的影响与工作满意度的间接影响，最终就是如图中三类JS间接影响因素+OCEAN的直接影响因素共同决定了用户的CPB倾向；而CPB倾向的高低又导致了内部破坏行为与轻度的迟到早退（缺勤发展成内部破坏），最终在实施内部破坏攻击后离职；

因此，依据上述动力模型，对于事后分析检测场景而言，基本步骤如下：

1. 将年度窗口用户中在职与离职用户分类，离职用户作为高CPB的潜在用户；
2. 将在职用户依据提取的ATF特征进行自动KMeans（2-10）聚类，筛选出CPB倾向较低的群簇，训练OCSVM；
3. 使用训练的OCSVM预测离职的测试用户，自动遍历参数选取最高Recall参数，输出结果；

上述工作还需要进行以下几个实验：

1. 计算ATF时使用的Leave Users还是LaidOff Users计算LCE特征；
2. 根据上述分析，OCEAN数据与JS或者CPB是重复特征，可以去掉不用；否则后续计算群簇公式时较难筛选；

首先使用去掉了OCEAN的新特征CPB\_ATF进行实验，后续全部步骤一致有：

此时的ATF文件为CERT5.2\_Static\_CPB\_ATF-0.1.csv

**PCA=1 Leave\_ATF**

**[0.05, 1.0, 0.8695652173913043, 0.8695652173913043, 0.0832016850974197, 0.10905452726363181, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',**

**max\_iter=-1, nu=0.05, random\_state=10, shrinking=True, tol=0.01,**

**verbose=False), 0.7931034482758621, 0.9, 1.0,]**

结果同使用OCEAN一样，也可以理解，因为OCEAN与后续的JS、CPB都相关，因而结果无影响。

DF函数结果保存在CERT5.2\_KMeans\_OCSVM\_CPB\_ATF\_Predictor\_Risk-0.1.csv

中

如果采用LaidOff来计算ATF，重新计算整个过程：结果略微比Leave ATF差，放弃！

[0.05, 1.0, 0.8405797101449275, 0.8405797101449275, 0.08109531332280147, 0.10605302651325663, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.05, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.7586206896551724, 0.8666666666666667, 1.0,

接下来，开始实验如何显示计算CPB来筛选群簇：

直接计算两个群簇的各个变量的均值/中位数，然后分别输入一个群簇筛选的标记计算：

我们先使用第一种计算公式看看：

上述计算中的结果来自于MinMax后的CPB\_ATF中的群簇中心值

计算

0 197

1 1564

0 ['2.5541490577132886', '3.4844853898564727']

1 ['2.6281383448135647', '3.622423493187438']

0 1564

1 197

0 ['2.5515736824051607', '3.5109122722209904']

1 ['2.4859452312892065', '3.3809321436702398']

上述结果竟然大群集的结果更高，不符合我们的预期

**0 1564**

**1 197**

**0 ['3.2753099575151476', '3.2186032810077254']**

**1 ['3.3508221785941816', '3.2531961210848963']**

即当放弃了缺勤率的系数之后，得到的大群集的CPB低于小群集

LED：0 ['0.7790327375125565', '1.0908185836192104']

LED：1 ['0.7418911236680935', '1.0392647777234978']

**0 1564**

**1 197**

**0 ['4.054342695027704', '4.309421864626936']**

**1 ['4.0927133022622755', '4.292460898808394']**

**如果改成CPB+缺勤比例和，结果也是大群簇CPB低**

二〇一九年一月二十二日星期二

上午12时46分

我们尝试对于CERT5.2的2000个用户进行分析，其中RandomForest计算特征重要性

user\_id,CPB-I,CPB-O,JS\_Score,Team\_CPB-I-mean,Team\_CPB-O-mean,Users-less-mean-A,Users-less-mean-A and C,Users-less-mean-C,Users-High-mean-N,Team\_CPB-I-median,Team\_CPB-O-median,leader-CPB-I,leader-CPB-O,dis\_ocean,avg\_dis\_ocean,dis\_os,avg\_dis\_os,email\_ratio,cnt\_send/recv,cnt\_s/r\_size,cnt\_s/r\_attach,cnt\_s/r\_days,cnt\_email\_days,cnt\_late\_days,cnt\_early\_days,month\_work\_days,

0 ['month\_work\_days', '0.17884890916218857']

1 ['dis\_ocean', '0.07411619038778294']

2 **['avg\_dis\_os', '0.043433668059703646']**

**3 ['cnt\_s\_size', '0.0378151194206492']**

**4 ['cnt\_r\_size', '0.03509465321522913']**

**5 ['cnt\_r\_days', '0.031086885168316883']**

**6 ['cnt\_late\_days', '0.030881186807124136']**

**7 [' cnt\_s\_days', '0.03038892187963833']**

**8 ['cnt\_email\_days', '0.030142745358936292']**

**9 ['cnt\_send', '0.0293265026667047']**

**10 ['leader-CPB-O', '0.029050196498913845']**

11 ['avg\_dis\_ocean', '0.027977327248848872']

**12 ['CPB-I', '0.027464482696853254']**

13 ['cnt\_s\_attach', '0.02729679136554958']

14 ['cnt\_r\_attach', '0.02665011531869494']

15 ['email\_ratio', '0.026266516707788606']

16 ['JS\_Score', '0.025379674583343538']

17 ['leader-CPB-I', '0.025039530053867755']

18 ['Users-less-mean-C', '0.02450696663068064']

19 ['cnt\_recv', '0.02395166416432676']

20 ['Team\_CPB-O-mean', '0.023851997210112337']

**21 ['CPB-O', '0.02332232207631709']**

22 ['cnt\_early\_days', '0.023163489332504086']

23 ['dis\_os', '0.02289069855011692']

24 ['Users-High-mean-N', '0.02284570236146889']

25 ['Team\_CPB-I-mean', '0.02177110145588932']

26 ['Team\_CPB-I-median', '0.02159279358901502']

27 ['Team\_CPB-O-median', '0.02108069837927249']

28 ['Users-less-mean-A and C', '0.019605549967053007']

29 ['Users-less-mean-A', '0.015157599683109677']

**二〇一九年一月二十四日星期四**

下午8时16分

**在Y老师看过一遍文章后，自己终于拿到了签字，并网上投出开始走流程。**

**今天，重新梳理下自己实验分析的CPB特征，并将其赋予正确的中心点CPB计算公式。**

**注意，默认使用OCEAN全部计算CPB的后缀0.1的ATF文件**

用户与LCE的OCEAN相关特征：user\_p

User\_p = user, o,c,e,a,c, cpb-i, cpb-o, dis\_ocean, avg\_dis\_ocean

原始ATF特征顺序为

User\_p, dis\_ldap, avg\_dis\_ldap, user\_led\_feat, user\_lc\_feat(email\_ratio+8+1)/cnt\_lc

突然发现了CERT4.2与CERT5.2计算时的疏忽，

CERT5.2计算dis\_ocean与dis\_ldap时，都使用公式进行了转换，变成了倒数，使得其与

**首先运行V09\_JS\_Predictor\_Main\_v01.py生成每月的均值化的LCE特征，然后与JS特征结合得到ATF特征即可；然后得到Static\_ATF特征，全部使用OCEAN决定的CPB/JS计算，全部使用Leave计算。**

**CERT4.2由于使用了直接的dis\_ocean/dis\_os，因此，需要进一步验证如果修改为CERT5.2的计算方式有无影响。**

**另外，需要重新计算Static/Dynamic，以及检测Leave与检测中高CPB的用户。**

**CERT5.2使用的距离映射CPB公式为**

**为了将CERT5.2与CERT4.2统一，以CERT4.2为标准，即：**

1. **直接计算dis\_ocean/os的结果特征标记为0.1**
2. **使用上述公式计算CPB影响的结果标记为0.2**

**二〇一九年一月二十五日星期五**

上午9时13分

今天上午继续实验，争取完成所有对比实验，给主观检测一个交代。

实验计划：

1. CERT5.2的静态实验（检测Leave与检测Leave+中高CPB）、动态实验（检测Leave）;
2. CERT4.2的静态实验（检测Leave与检测Leave+中高CPB）、动态实验（检测Leave）；

实验结果一：CERT5.2静态实验

* 检测Leave Users

**0 1564**

**1 197**

**中心选择：**

**计算后的中心CPB数值为：（CPBmean, CPBmedian），选择均值最低的即可；**

**0 ['4.048149096007543', '4.310527897331111']**

**1 ['4.096128204679356', '4.2970279816293715']**

**最优结果：**

**[0.0055, 1.0, 0.8985507246376812, 0.8985507246376812, 0.07424960505529225, 0.10155077538769385, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',max\_iter=-1, nu=0.0055, random\_state=10, shrinking=True, tol=0.01,verbose=False), 0.8620689655172413, 0.9, 1.0, array**

**其他参数结果**

*[0.0085, 1.0, 0.8695652173913043, 0.8695652173913043, 0.07740916271721959, 0.10355177588794397, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',max\_iter=-1, nu=0.0085, random\_state=10, shrinking=True, tol=0.01,*

*verbose=False), 0.8275862068965517, 0.8666666666666667, 1.0, array*

*[0.001, 1.0, 0.855072463768116, 0.855072463768116, 0.07266982622432859, 0.09854927463731866, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',*

*max\_iter=-1, nu=0.001, random\_state=10, shrinking=True, tol=0.01,*

*verbose=False), 0.8275862068965517, 0.8333333333333334, 1.0, array*

* **检测Leave Users + 中高CPB用户（Recall未变，但是FPR提升了一倍）**

[**0.0055**, 1.0, 0.8985507246376812, 0.8985507246376812, **0.16113744075829384**, 0.18409204602301152, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',max\_iter=-1, nu=0.0055, random\_state=10, shrinking=True, tol=0.01,verbose=False), **0.8620689655172413, 0.9, 1.0**, array

实验结果二：CERT5.2动态实验

采用CERT5.2的2010-06开始，到2011-05为止

2010-06：

[0.0005, 1.0, 1.0, **1.0（当月Recall的Insiders个数）, 1.0（当月Insider的Recall）, 16.0(当月FP个数), 0.0081799591002045（当月整体FPR）**, 0.008686765457332652, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **None, None, 1.0**, array([ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, 1,

1], dtype=int64)]

2010-07：

[0.0015, 1.0, 1.0, **4.0, 1.0, 7.0, 0.0036175710594315244**, 0.005673027333677153, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0015, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **1.0, None, 1.0**, array([ 1, 1, 1, -1, 1, 1, 1, 1, 1, 1, 1, 1, -1], dtype=int64)]

2010-08：

[0.005, 1.0, 0.875, **7.0, 0.875, 8.0, 0.004171011470281543**, 0.00778816199376947, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **1.0, 0.6666666666666666, 1.0**, array([ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, 1, 1, 1, 1, -1],

dtype=int64)]

2010-09：

[0.0005, 1.0, 1.0, **10.0, 1.0, 11.0, 0.005792522380200105**, 0.011000523834468309, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, None, array([ 1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,

1, 1, 1, 1, 1], dtype=int64)]

2010-10：

[0.0005, 1.0, 1.0, **7.0, 1.0, 7.0, 0.003723404255319149**, 0.007419183889772125, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2010-11：

[0.0055, 1.0, 0.8333333333333334, 5.0, 0.8333333333333334, 9.0, 0.0048205677557579, 0.007474639615589963, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0055, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 0.0, 1.0, array([ 1, 1, 1, 1, 1, 1, -1, 1, -1, 1, 1, 1, 1, 1, -1, -1, 1,

1. , dtype=int64)]

2010-12：

[0.0035, 1.0, 0.875, **7.0, 0.875, 12.0, 0.006497022198159177**, 0.010242587601078167, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0035, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **1.0, 0.8333333333333334, 1.0**, array([ 1, 1, 1, 1, -1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1,

1, 1, 1, 1, 1], dtype=int64)]

2011-01：

[0.001, 1.0, 1.0, 3.0, 1.0, 2.0, 0.001092896174863388, 0.002727768685215494, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.001, random\_state=10, shrinking=True, tol=0.01,

verbose=False), None, 1.0, None, array([ 1, -1, -1, 1, 1, -1, -1, -1, -1, 1, 1, -1], dtype=int64)]

2011-02：

[0.0005, 1.0, 1.0, 7.0, 1.0, 9.0, 0.004961411245865491, 0.008786381109280615, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2011-03：

[0.0005, 1.0, 1.0, 6.0, 1.0, 8.0, 0.0044469149527515284, 0.00775623268698061, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, None, None, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2011-04：

[0.0015, 1.0, 1.0, 5.0, 1.0, 6.0, 0.0033594624860022394, 0.006141820212171971, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0015, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([ 1, 1, -1, 1, 1, -1, 1, 1, 1, -1, 1, 1, 1, 1, -1],

dtype=int64)]

2011-05：

[0.0015, 1.0, 1.0, **4.0, 1.0, 8.0, 0.004514672686230248**, 0.006756756756756757, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0015, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([ 1, 1, 1, 1, 1, -1, 1, 1, 1, 1, -1, 1, 1, -1, 1],

dtype=int64)]

**CERT5.2数据集分析暂时告一段落！！**

——————————————————————————————————————————————————————————————————————————————

二〇一九年一月二十五日星期五

下午3时53分

继续分析CERT4.2数据集

实验一：静态分析

* 检测Leave Users：

KMeans聚类选择

0 732

1 113

**0 ['3.107954051001196', '3.169247520769576'] 同CERT5.2一样，选择低于均值的即可**

1 ['3.123946384376345', '2.9993110712047955']

[0.0028, 1.0, 0.8714285714285714, **0.8714285714285714（Recall）, 0.08279569892473118(FPR\_Total), 0.138,** OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0028, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **0.9333333333333333, 0.8666666666666667, 0.7**,

[0.0029, 1.0, 0.8714285714285714, 0.8714285714285714, 0.08172043010752689, 0.137, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0029, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.9333333333333333, 0.8666666666666667, 0.7,

[0.0022, 1.0, 0.8571428571428571, 0.8571428571428571, 0.06989247311827956, 0.125, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0022, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.9666666666666667, 0.7666666666666667, 0.8,

* 检测LeaveUsers+KMeans排除用户

[0.0028, 1.0, 0.8714285714285714, **0.8714285714285714, 0.2, 0.247**, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0028, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **0.9333333333333333, 0.8666666666666667, 0.7**, 召回率不变，但是FPR上升到了额20%（8%-->20%）

实验二：CERT4.2的动态逐月检测

其逐月分析时间2010-06:2011-04

[0.0048, 1.0, 0.5, **1.0, 0.5, 0.0, 0.00205761316872428**, 0.003080082135523614, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), '-1', '-2', 0.5, array([-1, -1, 1, 1, 1, -1, -1], dtype=int64)]

2010-07:

[0.0048, 1.0, 0.0, 0.0, 0.0, 0.0, 0.0031120331950207467, 0.0031023784901758012, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.0, '-2', 0.0, array([-1, 1, -1, 1, -1, 1, -1], dtype=int64)]

2010-08:

[0.0048, 1.0, 0.9090909090909091, 10.0, 0.9090909090909091, 0.0, 0.002107481559536354, 0.0125, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 0.75, 1.0, array([ 1, 1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, 1],

dtype=int64)]

2010-09:

[0.0048, 1.0, 0.875, 7.0, 0.875, 0.0, 0.005330490405117271, 0.012684989429175475, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 0.8571428571428571, '-3', array([ 1, 1, 1, 1, 1, 1, 1, -1, 1, 1, 1, 1, 1], dtype=int64)]

2010-10:

[0.0048, 1.0, 0.5833333333333334, 7.0, 0.5833333333333334, 0.0, 0.004343105320304018, 0.011789924973204717, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.7142857142857143, 0.6666666666666666, 0.0, array([ 1, -1, 1, 1, 1, 1, -1, 1, 1, 1, 1, -1, -1, -1, 1, 1],

dtype=int64)]

2010-11:

[0.0048, 1.0, 0.75, 6.0, 0.75, 0.0, 0.0033003300330033004, 0.009814612868047983, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.8, 0.5, 1.0, array([ 1, -1, 1, 1, 1, 1, 1, 1, -1, 1, -1, 1], dtype=int64)]

2010-12:

[0.0048, 1.0, 1.0, 8.0, 1.0, 0.0, 0.008918617614269788, 0.017679558011049725, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([ 1, 1, 1, 1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],

dtype=int64)]

2011-01:

[0.0048, 1.0, 0.3333333333333333, 1.0, 0.3333333333333333, 0.0, 0.006779661016949152, 0.007882882882882882, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 0.0, '-3', array([ 1, 1, -1, -1, 1, 1, 1, 1, -1, 1, -1], dtype=int64)]

2011-02:

[0.0048, 1.0, 0.8333333333333334, 5.0, 0.8333333333333334, 0.0, 0.003444316877152698, 0.009122006841505131, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 0.5, '-3', array([ 1, 1, -1, 1, 1, 1, 1, 1, -1, 1], dtype=int64)]

2011-03:

[0.0048, 1.0, 0.75, 3.0, 0.75, 0.0, 0.006952491309385863, 0.010380622837370242, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.5, 1.0, '-3', array([ 1, -1, 1, 1, 1, 1, 1, 1, 1, 1, -1], dtype=int64)]

2011-04:

[0.0048, 1.0, 0.6, **3.0, 0.6, 0.0, 0.005875440658049354**, 0.009345794392523364, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), '-1', 0.5, 1.0, array([ 1, 1, -1, 1, 1, -1, -1, 1, 1, 1, 1], dtype=int64)]

目前分析的ATF特征去掉了用户的OCEAN特征，而是CERT5.2\_Static\_CPB\_ATF-0.1.csv

具体包含的特征为：

user\_id,

CPB-I,CPB-O,(2)

JS\_Score,(1)

Team\_CPB-I-mean,Team\_CPB-O-mean,Users-less-mean-A,Users-less-mean-A and C,

Users-less-mean-C,Users-High-mean-N,Team\_CPB-I-median,Team\_CPB-O-median, (8)

leader-CPB-I,leader-CPB-O, (2)

dis\_ocean,avg\_dis\_ocean,dis\_os,avg\_dis\_os, (4)

email\_ratio, (1)

cnt\_send，cnt\_recv, (2)

cnt\_s\_size, cnt\_r\_size, (2)

cnt\_s\_attach,cnt\_r\_attach (2)

cnt\_s\_days, cnt\_r\_days (2)

cnt\_email\_days, (1)

cnt\_late\_days,cnt\_early\_days,month\_work\_days, (3)

上述原始特征一共30个维度

我们选择一个用户的特征实地看一下

MMK1532,

-6.56,-9.51,

6.76,

-13.761999999999999,-24.254666666666665,16,7,

12,16,-14.555000000000001,-25.135,

-4.79,-10.23,

**12.3121009712,0.286327929563,0.245534276386,0.00571009945084,**

-0.627906976744,

0.186046511628,0.860465116279,

5859.51162791,208139.372093,

0.0,0.186046511628,

0.186046511628,0.813953488372,

0.953488372093,

169.0,161.0,356.0,

**二〇一九年一月二十八日星期一**

**下午2时29分**

今天在回顾整个检测过程中，发现了一点小瑕疵：

对于在职用户采用KMeans筛选低CPB用户时，主要方法是对在职用户进行了MinMax后进行聚类后中心点计算；

而最终训练OCSVM检测离职用户时，则是对所有用户进行了PCA，然后进行scale后训练OCSVM检测离职用户，这其中选择过程中在职用户处理方式不同，不能简单用筛选的索引作为训练依据

做个简单的修正：

1. 对所有CPB特征做MinMax；
2. 对在职用户的CPB特征进行自动聚类，选择中低CPB用户索引；这一过程中不再针对在职用户归一化；
3. 映射回MinMax后的CPB特征，并统一做PCA后训练OCSVM测试Lever，结果得到最优：

[0.0015, 1.0, 1.0, **1.0, 0.0889942074776198**, 0.11905952976488245, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0015, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **1.0, 1.0, 1.0**, array

如果不做PCA直接运算呢？

**如果做自动PCA，则有：**

**[0.003, 1.0, 1.0, 1.0, 0.05897840968931016, 0.09054527263631816, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',**

**max\_iter=-1, nu=0.003, random\_state=10, shrinking=True, tol=0.01,**

**verbose=False), 1.0, 1.0, 1.0, array**

**即一次事前的MinMax，然后一次训练测试前的自动PCA，则有自动PCA=30维度；**

**如果是静态模式下检测HM用户，则有**

**[0.003, 1.0, 1.0, 1.0, 0.1579778830963665, 0.18459229614807404, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',**

**max\_iter=-1, nu=0.003, random\_state=10, shrinking=True, tol=0.01,**

**verbose=False), 1.0, 1.0, 1.0, array**

其结果定性同之前的一样。

这次结果比之前要好，全部检出，而且FP还低

1. 自动PCA=30
2. Reall = 100%

下面开始逐月分析

2010-06：

[0.0005, 1.0, 1.0, 1.0, 1.0, 17.0, 0.008691206543967281, 0.009197751660705161, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), None, None, 1.0, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2010-07：

[0.0005, 1.0, 1.0, 4.0, 1.0, 9.0, 0.004651162790697674, 0.006704486848891181, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, None, 1.0, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2010-08：

[0.0005, 1.0, 1.0, 8.0, 1.0, 8.0, 0.004171011470281543, 0.008307372793354102, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([ 1, 1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],

dtype=int64)]

2010-09：

[0.0005, 1.0, 1.0, 10.0, 1.0, 12.0, 0.00631911532385466, 0.011524358302776323, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, None, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],

dtype=int64)]

2010-10：

[0.0005, 1.0, 1.0, 7.0, 1.0, 7.0, 0.003723404255319149, 0.007419183889772125, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2010-11：

[0.0005, 1.0, 1.0, 6.0, 1.0, 12.0, 0.006427423674343867, 0.009610250934329953, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2010-12：

[0.0005, 1.0, 1.0, 8.0, 1.0, 14.0, 0.007579859231185706, 0.011859838274932614, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1],

dtype=int64)]

2011-01：

[0.0005, 1.0, 1.0, 3.0, 1.0, 9.0, 0.004918032786885246, 0.006546644844517185, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), None, 1.0, None, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2011-02：

[0.0005, 1.0, 1.0, 7.0, 1.0, 9.0, 0.004961411245865491, 0.008786381109280615, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2011-03：

[0.0005, 1.0, 1.0, 6.0, 1.0, 8.0, 0.0044469149527515284, 0.00775623268698061, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, None, None, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2011-04：

[0.0005, 1.0, 1.0, 5.0, 1.0, 10.0, 0.005599104143337066, 0.008375209380234505, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2011-05：

[0.0005, 1.0, 1.0, 4.0, 1.0, 10.0, 0.0056433408577878106, 0.007882882882882882, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0005, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 1.0, array([ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, 1],

dtype=int64)]

接下来进行CERT5.2的特征重要性分析：

对原始staticATF做MinMax，然后使用RandomForest分析重要性

0 ['month\_work\_days', '0.21636456963661355']

1 ['avg\_dis\_os', '0.1366945319822552']

2 ['dis\_os', '0.09491250416410814']

3 ['dis\_ocean', '0.07636647904599639']

4 ['cnt\_r\_size', '0.03758211347051841']

5 ['cnt\_r\_days', '0.03543410201921143']

6 ['cnt\_s\_days', '0.028562336493579022']

7 ['cnt\_send', '0.024473871074277436']

8 ['cnt\_late\_days', '0.023263542189561538']

9 ['email\_ratio', '0.022933750116291653']

10 ['cnt\_r\_attach', '0.022866616186046834']

11 ['avg\_dis\_ocean', '0.02213623135109763']

12 ['cnt\_s\_size', '0.02140403091583166']

13 ['cnt\_early\_days', '0.02090234974867144']

14 ['cnt\_recv', '0.0206503203239596']

15 ['CPB-I', '0.02003440167661133']

16 ['leader-CPB-O', '0.01804581237256772']

17 ['JS\_Score', '0.017166318123029174']

18 ['CPB-O', '0.016745102480930113']

19 ['cnt\_s\_attach', '0.016305271576163387']

20 ['Team\_CPB-O-mean', '0.014280258639572309']

21 ['Team\_CPB-I-mean', '0.01352476068069222']

22 ['Users-High-mean-N', '0.012861661570920828']

23 ['leader-CPB-I', '0.012850377143345717']

24 ['Users-less-mean-C', '0.012765287051208357']

25 ['Team\_CPB-O-median', '0.012010201822129156']

26 ['Team\_CPB-I-median', '0.010831518948500278']

27 ['Users-less-mean-A and C', '0.009061149421312163']

28 ['Users-less-mean-A', '0.008970529774996973']

我们按照特征类别进行分析，可以得到：

0 ['WorkTime', '0.17254568114720192']

1 ['Leave\_Contacts', '0.1717110343811845']

2 ['CPB-O', '0.16475842180082417']

3 ['JS\_Score', '0.16430638151841218']

4 ['CPB-I', '0.1636015582032164']

5 ['Work\_CPBs', '0.1630769229491593']

二〇一九年一月二十八日星期一

下午11时51分

接下来，我们试着对CERT5.2进行少许的案例分析（采用分组特征）

1. CPB\_FeatsLM的均值特征
2. Insiders\_1
3. Insiders\_2
4. Insiders\_3

计算得到CERT5.2中群簇0（1564个）的均值特征为

Cluster 1

[0.4921393 0.51142666 0.50344819 1. 1. 0.

0. 0. 0. 1. 1. 0.54746193

0.47281222 0.35884406 0.46948578 0.22324165 0.08237508 0.63861849

0.14134065 0.10025865 0.26597656 0.17411393 0.11054657 0.11821991

0.21782442 0.16584112 0.19081457 0.21499769 0.36105584 0.0477678 ]

Cluseter 0 (1564)

[0.48910005 0.51125164 0.50229733 0.17806591 0.20199699 0.5781269

0.46776428 0.55608838 0.59100597 0.25762216 0.18817143 0.53839514

0.44814785 0.27897101 0.49302197 0.19853182 0.10848115 0.51437663

0.13276488 0.12528375 0.23784272 0.24771765 0.11956776 0.142736

0.25928251 0.27493893 0.31256463 0.22452337 0.38475448 0.0574136 ]

结果是

**[0.48910005, 0.51125164, 0.50229733, 4.00538501, 3.133516780000001, 0.92184248]**

**['0.7907037133212153', '0.7758957654723125', '0.19323483670295488', '5.168403401197614', '1.937963028196464', '0.582554112554303']**

**['0.6517787587639574', '0.6067318132464712', '0.34525660964230176', '5.941097744900533', '3.2096073838277905', '0.3135802469137778']**

**['0.5481693066735913', '0.4397394136807816', '0.5458786936236392', '3.6817205832000797', '3.9743453263491824', '0.7083203969517576']**

需要计算的是分组别特征

紧接着我们针对CERT4.2数据集进行实验

首先同样是静态模式：

[0.0021, 1.0, 0.9142857142857143, **0.9142857142857143, 0.06989247311827956**, 0.129, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0021, random\_state=10, shrinking=True, tol=0.01,

verbose=False), **0.9333333333333333, 0.9, 0.9**, array

然后是检测中高CPB用户：

[0.0021, 1.0, 0.9142857142857143, 0.9142857142857143, **0.1870967741935484**, 0.238, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0021, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.9333333333333333, 0.9, 0.9, array

经过分析，Static与Dynamic相比Recall一致，但是Static误报有65个，而Dynamic只有54个，因而Dynamic的FP降低了14%

2010-06：

[0.0048, 1.0, 0.5, **1.0, 0.5, 4.0 0.00411522633744856,** 0.00513347022587269, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), '-1', '-2', 0.5, array([ 1, -1, 1, 1, -1, 1, 1], dtype=int64)]

2010-07：

[0.0048, 1.0, 1.0, 3.0, 1.0, 3.0, 0.0031120331950207467, 0.0062047569803516025, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, '-2', 1.0, array([ 1, 1, 1, 1, 1, -1, 1], dtype=int64)]

2010-08：

[0.0048, 1.0, 0.9090909090909091, 10.0, 0.9090909090909091, 3.0, 0.003161222339304531, 0.013541666666666667, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.8333333333333334, 1.0, 1.0, array([ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1],

dtype=int64)]

2010-09：

[0.0048, 1.0, 1.0, 8.0, 1.0, 4.0, 0.0042643923240938165, 0.012684989429175475, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, '-3', array([ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, 1], dtype=int64)]

2010-10：

[0.0048, 1.0, 0.8333333333333334, 10.0, 0.8333333333333334, 4.0, 0.004343105320304018, 0.015005359056806002, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 1.0, 0.0, array([ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, 1, -1, 1, 1],

dtype=int64)]

2010-11：

[0.0048, 1.0, 0.875, 7.0, 0.875, 4.0, 0.0044004400440044, 0.011995637949836423, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 0.5, 1.0, array([ 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, -1, 1], dtype=int64)]

2010-12：

[0.0048, 1.0, 0.875, 7.0, 0.875, 9.0, 0.010033444816053512, 0.017679558011049725, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 0.75, 1.0, array([ 1, 1, 1, 1, 1, 1, 1, 1, -1, 1, 1, 1, 1, 1, 1, 1, 1],

dtype=int64)]

2011-01：

[0.0048, 1.0, 0.6666666666666666, 2.0, 0.6666666666666666, 8.0, 0.00903954802259887, 0.01126126126126126, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 1.0, 0.5, '-3', array([ 1, 1, 1, -1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2011-02：

[0.0048, 1.0, 0.8333333333333334, 5.0, 0.8333333333333334, 3.0, 0.003444316877152698, 0.009122006841505131, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.75, 1.0, '-3', array([ 1, 1, 1, 1, 1, -1, 1, 1, -1, 1], dtype=int64)]

2011-03：

[0.0048, 1.0, 0.75, 3.0, 0.75, 7.0, 0.008111239860950173, 0.011534025374855825, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), 0.5, 1.0, '-3', array([-1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]

2011-04：

[0.0048, 1.0, 1.0, 5.0, 1.0, 5.0, 0.005875440658049354, 0.011682242990654205, OneClassSVM(cache\_size=200, coef0=0.0, degree=3, gamma='auto', kernel='rbf',

max\_iter=-1, nu=0.0048, random\_state=10, shrinking=True, tol=0.01,

verbose=False), '-1', 1.0, 1.0, array([ 1, -1, 1, 1, 1, 1, 1, 1, 1, 1, 1], dtype=int64)]