Hi David,

Please find the clustering result using k-mean algorithm in kmean\_result.doc

Here are steps how I preprocess and cluster data:

**0. Transform features**

* IP -> country
* For “RevokePush”, there are 4 new attributes

1. RevokePushTF: True if RevokePush is not missing (original value is in datetime)

Calculate RevokeDays = RevokePush – FirstUseDate, then

1. RevokePushBafore = 1 if RevokeDays <0, = 0 otherwise
2. RevokePushSameDay = 1 if RevokeDays = 0 , = 0 otherwise
3. RevokePushAfter = RevokeDays if RevokeDays>0, = 0 otherwise

* For "BlockPush" and "Uninstalled", there are 3 new attributes for each one (original value is in datetime)

1. BlockPushTF, UninstalledTF: True if BlockPush / Uninstalled is not missing

Calculate BlockPushDays = BlockPush - FirstUseDate,

UninstalledDays = Uninstalled - FirstUseDate then

1. BlockPushSameday, UninstalledSameday = 1 if Days = 0,= 0 otherwise
2. BlockPushAfter, UninstalledAfter = 1 if Days > 0,= 0 otherwise

* Timezone -> Timezone/60

**1. Normalize features**

- Integer features: convert to [0,1] using min\_max scaler

Read more about this function at http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.MinMaxScaler.html

- Categorical features: one-2- encoder

Read more at http://statsmodels.sourceforge.net/ipdirective/generated/scikits.statsmodels.tools.tools.categorical.html

- Boolean features: keep original values

**2. Impute missing value:** replace missing value with mean for integer features and replace with mode for boolean and categorical features

Read more at http://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Imputer.html

**3. Select features**

3.1 Remove features with low and zero variance: "AdOptedIn","AllowiBeacon","HaveUniqueGlobalID","OS","SignIn","EmailExist","UserType"

3.2 Perform PCA to select the best features:

The selected features are

AllowPush

AllowGeo

AllowFeaturePush

AllowBT

RevokePushBefore

RevokePushAfterDays

UninstalledTF

UninstalledSameday

UninstalledAfter

NumCampaignMatch

NumCrash

InstallDays

PushCount

CorrectQuestion

LastUpdateDays

ScreenHeight

Timezone

OSVersion

DeviceModel

ScreenWidth

Read more about PCA at http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html

**4. Cluster data using k-mean algorithm and select the best k**

Use Silhouette coefficients (the bigger the better) to select the best number of clusters from k-mean algorithm

Read more at http://scikit-learn.org/stable/auto\_examples/cluster/plot\_kmeans\_silhouette\_analysis.html

Here are a value of k and its silhouette coefficients

(k, silhoette coefficients)

(3, 0.15419767227232772)

(4, 0.1749548768057555) \* best k=4

(5, 0.16251883723896135)

(6, 0.16043009203494993)

(7, 0.16346565100465482)

(8, 0.12729074920578809)

(9, 0.13692302565338529)

(10, 0.12924594523964311)

(11, 0.12676767327019539)

(12, 0.1241204428781716)

(13, 0.12448054715003889)

(14, 0.1238881926277654)

(15, 0.1209973068278709)

(16, 0.12095143965823443)

(17, 0.12024508046799569)

(18, 0.11391827979012258)

(19, 0.11109333514948599)

Here are frequency distributions of features

For boolean features: 1 = True, 0 = False

**Each feature shows frequency of each feature value**

(0 = False, 1 = True)

1: AllowPush

(' 0', 577)

(' 1', 551)

sum = 1128

2: AdOptedIn

(' 0', 1128)

sum = 1128

3: NumCampaignMatch

(' 10', 5)

(' 11', 5)

(' 12', 67)

(' 13', 3)

(' 16', 17)

(' 2', 66)

(' 3', 25)

(' 0', 628)

(' 1', 76)

(' 6', 19)

(' 7', 26)

(' 4', 53)

(' 5', 20)

(' 8', 100)

(' 9', 18)

sum = 1128

4: Carrier

(' TRUE-H', 10)

(' Orange Armenia', 1)

(' VODAFONE TR', 1)

(' Telstra', 5)

(' VinaPhone', 1)

(' SMC HK', 5)

(' T-Mobile A', 1)

(' Etisalat', 3)

(' AT&T', 89)

(' STC', 5)

(' Oi', 3)

(' StarHub', 13)

(' NetCom', 1)

(' non-ascii8', 2)

(' T-Mobile NL', 1)

(' Virgin Mobile', 3)

(' vf es', 2)

(' E-Plus', 2)

(' non-ascii2', 4)

(' non-ascii3', 11)

(' non-ascii1', 12)

(' non-ascii6', 3)

(' dtac', 4)

(' non-ascii4', 2)

(' non-ascii5', 2)

(' Zain KSA', 1)

(' OPTUS', 10)

(' Turkcell', 1)

(' 3', 9)

(' TELUS', 11)

(' NOS', 1)

(' Swisscom', 3)

(' Vodacom', 24)

(' Vodafone India', 36)

(' cricket', 2)

(' Reliance', 2)

(' vodafone P', 1)

(' Bouygues Telecom', 8)

(' DiGi', 1)

(' Vodafone.de', 7)

(' O2', 14)

(' Antel', 1)

(' mt:s', 2)

(' Orange Romania', 1)

(' VIVA', 1)

(' MobiCom', 1)

(' movistar', 2)

(' Claro Brasil', 4)

(' Zain SA', 1)

(' TIGO HND', 2)

(' BTC', 1)

(' TELEKOM.RO', 1)

(' Zain JO', 7)

(' Indosat', 1)

(' Telenor', 2)

(' CSL', 8)

(' TelkomSA', 4)

(' KPN NL', 2)

(' KDDI', 2)

(' Claro Dominicana', 1)

(' Mobile TeleSystems', 2)

(' Vodafone', 1)

(' Beeline', 3)

(' touch', 1)

(' CLARO PERU', 1)

(' Kcell', 2)

(' Viettel', 1)

(' Maxis', 3)

(' Sunrise', 10)

(' Sprint', 24)

(' Zong', 6)

(' XL', 2)

(' csl.', 1)

(' Zain BH', 1)

(' Yoigo', 1)

(' vf nl', 2)

(' TIM', 3)

(' Meteor', 1)

(' Mobilink', 12)

(' VodaCom', 1)

(' ENTEL PCS', 1)

(' Ooredoo', 5)

(' Verizon', 66)

(' MegaFon', 1)

(' M1 Singapore', 9)

(' Telenor PK', 8)

(' CELCOM', 1)

(' KT', 2)

(' Warid', 12)

(' non-ascii7', 1)

(' Orange Dominicana', 1)

(' SingTel', 2)

(' U.S. Cellular', 1)

(' MTS', 1)

(' MTN', 17)

(' Mobily', 2)

(' Carrier', 62)

(' Vodafone IT', 1)

(' iPad', 1)

(' Bluegrass', 1)

(' SFR', 7)

(' Mobinil', 3)

(' T-Mobile', 47)

(' OMANTEL', 1)

(' Cellcom', 1)

(' Zain KW', 2)

(' Telkomsel', 2)

(' Lebara', 2)

(' VIVO', 9)

(' CMHK', 10)

(' voda AU', 18)

(' AIS', 14)

(' IUSACELL', 1)

(' Salt', 1)

(' bmobile', 1)

(' giffgaff', 9)

(' BASE', 1)

(' Orange', 2)

(' du', 9)

(' DSTCom', 1)

(' LIME', 1)

(' Singtel', 21)

(' Play', 1)

(' Orange France', 4)

(' tbaytel', 1)

(' Koodo', 3)

(' Orange Espa\\xc3\\xb1a', 1)

(' Free', 5)

(' AirTel', 27)

(' vodafone UK', 14)

(' Rogers', 20)

(' vf ro', 1)

(' Proximus', 1)

(' Cell C', 3)

(' M\\xc3\\xa9ditel', 1)

(' Dialog', 6)

(' SMART', 7)

(' Telcel', 8)

(' Ufone', 34)

(' Bell', 12)

(' Nextel', 1)

(' Vodafone QA', 1)

(' Globe', 26)

(' AVEA', 2)

(' fido', 19)

(' SK Telecom', 1)

(' Telekom.de', 5)

(' etisalat', 24)

(' POST', 1)

(' EE', 19)

(' 2degrees', 1)

(' Three', 17)

(' Other', 50)

(' non-ascii9', 1)

(' 3 Ireland', 1)

(' AIRCEL', 1)

(' Videotron', 5)

(' o2-de', 7)

sum = 1128

Note: non-ascii is in non-English language

5: AppVersion

(' 1071256', 1003)

(' 1061241', 125)

sum = 1128

6: AllowiBeacon

(' 1', 1128)

sum = 1128

7: AllowGeo

(' 0', 667)

(' 1', 461)

sum = 1128

8: AllowFeaturePush

(' 0', 436)

(' 1', 692)

sum = 1128

9: ScreenHeight

(' 736', 133)

(' 480', 164)

(' 568', 439)

(' 667', 392)

sum = 1128

10: AllowBT

(' 1', 85)

(' None', 128)

(' 0', 915)

sum = 1128

11: HaveUniqueGlobalID

(' 1', 1128)

sum = 1128

12: NumCrash

(' 18', 1)

(' 3', 1)

(' 0', 1114)

(' 1', 10)

(' 2', 2)

sum = 1128

13: DailyUsage

(' 0', 3)

(' 1', 1125)

sum = 1128

14: Country

(' Australia', 41)

(' Indonesia', 5)

(' Poland', 1)

(' Singapore', 50)

(' South Africa', 50)

(' Guernsey', 1)

(' Bahrain', 1)

(' Algeria', 1)

(' Belgium', 2)

(' Brunei Darussalam', 1)

(' France- Metropolitan', 1)

(' United Kingdom', 77)

(' Tanzania- United Republic of', 1)

(' Lebanon', 1)

(' Jordan', 8)

(' Uganda', 1)

(' China', 29)

(' Thailand', 28)

(' New Zealand', 1)

(' Germany', 23)

(' Spain', 6)

(' Macau', 1)

(' Sri Lanka', 6)

(' Netherlands', 4)

(' Oman', 2)

(' Portugal', 2)

(' Philippines', 41)

(' Kazakhstan', 4)

(' Uruguay', 1)

(' Mauritius', 1)

(' Nigeria', 4)

(' Myanmar', 1)

(' Luxembourg', 1)

(' Asia/Pacific Region', 1)

(' Korea- Republic of', 3)

(' Vietnam', 6)

(' Turkey', 4)

(' Bulgaria', 1)

(' Switzerland', 16)

(' Pakistan', 96)

(' Saudi Arabia', 7)

(' Honduras', 2)

(' Mexico', 10)

(' Brazil', 19)

(' United Arab Emirates', 32)

(' Estonia', 1)

(' Egypt', 5)

(' Azerbaijan', 4)

(' Romania', 3)

(' Taiwan', 1)

(' India', 77)

(' Malaysia', 7)

(' Iran- Islamic Republic of', 4)

(' Argentina', 2)

(' Kenya', 1)

(' Russian Federation', 13)

(' Japan', 8)

(' Dominican Republic', 2)

(' Qatar', 3)

(' Austria', 1)

(' Hong Kong', 32)

(' Canada', 79)

(' Serbia', 2)

(' Armenia', 1)

(' Unknown', 1)

(' Italy', 1)

(' Kuwait', 3)

(' France', 23)

(' Mongolia', 2)

(' Nepal', 1)

(' United States', 246)

(' Morocco', 1)

(' Trinidad and Tobago', 1)

(' Peru', 1)

(' Ireland', 2)

(' Bahamas', 1)

(' Norway', 3)

(' Tunisia', 1)

(' Chile', 1)

sum = 1128

15: LastUpdateDays

(' 0', 619)

(' -1', 509)

sum = 1128

16: DeviceModel

(' Unknown iPad4', 5)

(' Unknown iPad5', 13)

(' iPad mini 1G', 5)

(' iPhone 5s', 201)

(' iPhone 4S', 52)

(' Unknown iPad2', 11)

(' iPad Air', 14)

(' Unknown iPhone8', 65)

(' iPad 4', 13)

(' iPod touch 5G', 2)

(' iPad 3', 5)

(' iPad 2', 1)

(' iPhone 5c', 30)

(' iPhone 6', 382)

(' iPhone 4', 38)

(' iPhone 5', 132)

(' iPhone 6 Plus', 152)

(' iPad mini 2G', 7)

sum = 1128

17: BlockPushTF

(' 0', 1070)

(' 1', 58)

sum = 1128

18: BlockPushSameday

(' 0', 1128)

sum = 1128

19: BlockPushAfterDays

(' 10', 2)

(' 11', 1)

(' 39', 2)

(' 14', 3)

(' 15', 2)

(' 16', 5)

(' 17', 5)

(' 18', 2)

(' 33', 1)

(' 57', 2)

(' 13', 2)

(' 61', 1)

(' 54', 1)

(' 24', 1)

(' 26', 1)

(' 23', 3)

(' 29', 1)

(' 28', 2)

(' 41', 1)

(' 2', 1)

(' 0', 1070)

(' 1', 2)

(' 6', 2)

(' 7', 3)

(' 4', 4)

(' 5', 1)

(' 8', 2)

(' 9', 2)

(' 31', 2)

(' 70', 1)

sum = 1128

20: OS

(' ios', 1128)

sum = 1128

21: OSVersion

(' 840', 70)

(' 841', 75)

(' 901', 76)

(' 830', 73)

(' 902', 260)

(' 710', 2)

(' 711', 9)

(' 712', 59)

(' 706', 1)

(' 800', 5)

(' 704', 7)

(' 820', 17)

(' 802', 4)

(' 702', 3)

(' 910', 351)

(' 810', 15)

(' 813', 21)

(' 812', 18)

(' 811', 15)

(' 900', 47)

sum = 1128

22: RevokePushTF

(' 0', 498)

(' 1', 630)

sum = 1128

23: RevokePushBefore

(' 0', 593)

(' 1', 535)

sum = 1128

24: RevokePushSameday

(' 0', 1077)

(' 1', 51)

sum = 1128

25: RevokePushAfterDays

(' 10', 1)

(' 11', 2)

(' 12', 3)

(' 39', 1)

(' 14', 2)

(' 15', 1)

(' 35', 1)

(' 13', 2)

(' 25', 3)

(' 24', 1)

(' 26', 1)

(' 21', 3)

(' 20', 1)

(' 47', 1)

(' 44', 2)

(' 43', 1)

(' 40', 1)

(' 2', 1)

(' 3', 2)

(' 0', 1084)

(' 1', 4)

(' 6', 1)

(' 7', 1)

(' 4', 2)

(' 8', 2)

(' 9', 4)

sum = 1128

26: SignIn

(' 1', 1128)

sum = 1128

27: UninstalledTF

(' 0', 1070)

(' 1', 58)

sum = 1128

28: UninstalledSameday

(' 0', 1128)

sum = 1128

29: UninstalledAfter

(' 10', 2)

(' 11', 1)

(' 39', 2)

(' 14', 3)

(' 15', 2)

(' 16', 5)

(' 17', 5)

(' 18', 2)

(' 33', 1)

(' 57', 2)

(' 13', 2)

(' 61', 1)

(' 54', 1)

(' 24', 1)

(' 26', 1)

(' 23', 3)

(' 29', 1)

(' 28', 2)

(' 41', 1)

(' 2', 1)

(' 0', 1070)

(' 1', 2)

(' 6', 2)

(' 7', 3)

(' 4', 4)

(' 5', 1)

(' 8', 2)

(' 9', 2)

(' 31', 2)

(' 70', 1)

sum = 1128

30: ScreenWidth

(' 375', 392)

(' 320', 603)

(' 414', 133)

sum = 1128

31: EmailExist

(' 1', 1128)

sum = 1128

32: EmailAddress

(' 0', 255)

(' 1', 873)

sum = 1128

33: InstallDays

(' 10', 19)

(' 11', 12)

(' 12', 21)

(' 13', 19)

(' 14', 14)

(' 15', 12)

(' 16', 11)

(' 17', 14)

(' 18', 13)

(' 19', 16)

(' 30', 6)

(' 31', 6)

(' 36', 3)

(' 37', 4)

(' 34', 4)

(' 35', 8)

(' 57', 1)

(' 39', 2)

(' -1', 120)

(' 55', 2)

(' 58', 1)

(' 59', 1)

(' 49', 6)

(' 48', 6)

(' 63', 2)

(' 50', 3)

(' 32', 5)

(' 53', 3)

(' 66', 1)

(' 33', 5)

(' 69', 1)

(' 42', 1)

(' 56', 2)

(' 25', 11)

(' 24', 3)

(' 27', 10)

(' 26', 15)

(' 21', 10)

(' 20', 10)

(' 23', 3)

(' 22', 10)

(' 47', 4)

(' 43', 3)

(' 44', 3)

(' 29', 8)

(' 28', 5)

(' 41', 2)

(' 51', 5)

(' 2', 37)

(' 3', 36)

(' 0', 408)

(' 1', 51)

(' 6', 23)

(' 7', 24)

(' 4', 29)

(' 5', 22)

(' 8', 18)

(' 9', 21)

(' 46', 2)

(' 38', 3)

(' 52', 3)

(' 72', 1)

(' 40', 4)

sum = 1128

34: PushCount

(' 10', 5)

(' 11', 5)

(' 12', 67)

(' 13', 3)

(' 15', 1)

(' 16', 16)

(' 2', 66)

(' 3', 26)

(' 0', 628)

(' 1', 76)

(' 6', 17)

(' 7', 27)

(' 4', 52)

(' 5', 22)

(' 8', 99)

(' 9', 18)

sum = 1128

35: Timezone

(' 10', 10)

(' 11', 29)

(' 13', 1)

(' -2', 14)

(' 3', 45)

(' 2', 106)

(' -3', 8)

(' -4', 109)

(' -6', 40)

(' -7', 46)

(' -5', 115)

(' 0', 42)

(' 1', 89)

(' 6', 3)

(' 7', 38)

(' 4', 41)

(' 5', 189)

(' -10', 1)

(' 8', 169)

(' 9', 11)

(' -8', 22)

sum = 1128

36: UserType

(' Paid', 1)

(' Free', 1127)

sum = 1128

37: Questions

(' 10', 599)

(' 39', 2)

(' 14', 1)

(' 15', 2)

(' 32', 1)

(' 30', 70)

(' 50', 29)

(' 51', 2)

(' 150', 3)

(' 60', 110)

(' 552', 1)

(' 136', 1)

(' 67', 1)

(' 134', 1)

(' 113', 2)

(' 110', 12)

(' 160', 2)

(' 65', 1)

(' 54', 1)

(' 115', 1)

(' 95', 2)

(' 109', 1)

(' 81', 1)

(' 80', 8)

(' 85', 1)

(' 84', 1)

(' 25', 2)

(' 20', 144)

(' 170', 3)

(' 36', 1)

(' 45', 1)

(' 66', 1)

(' 40', 51)

(' 105', 1)

(' 186', 1)

(' 180', 2)

(' 111', 1)

(' 206', 1)

(' 173', 1)

(' 200', 1)

(' 18', 1)

(' 209', 2)

(' 76', 2)

(' 240', 1)

(' 75', 2)

(' 123', 1)

(' 70', 12)

(' 120', 15)

(' 90', 15)

(' 106', 1)

(' 100', 7)

(' 96', 1)

(' 97', 1)

sum = 1128

38: CorrectQuestion

(' 10', 100)

(' 11', 2)

(' 12', 4)

(' 13', 4)

(' 14', 4)

(' 15', 9)

(' 16', 1)

(' 17', 9)

(' 54', 2)

(' 33', 11)

(' 30', 46)

(' 57', 16)

(' 50', 54)

(' 27', 7)

(' 52', 3)

(' 35', 17)

(' 32', 1)

(' 39', 3)

(' 44', 2)

(' 58', 5)

(' 29', 2)

(' 59', 2)

(' 9', 5)

(' 49', 5)

(' 37', 5)

(' 22', 2)

(' 61', 2)

(' 60', 50)

(' 63', 11)

(' 62', 5)

(' 65', 14)

(' 64', 3)

(' 67', 7)

(' 55', 18)

(' 68', 3)

(' 42', 4)

(' 80', 27)

(' 87', 1)

(' 56', 4)

(' 117', 1)

(' 25', 8)

(' 24', 3)

(' 46', 5)

(' 26', 4)

(' 21', 1)

(' 20', 74)

(' 23', 10)

(' 48', 8)

(' 47', 11)

(' 36', 4)

(' 45', 19)

(' 66', 4)

(' 43', 12)

(' 28', 6)

(' 41', 2)

(' 40', 63)

(' 2', 2)

(' 3', 3)

(' 0', 268)

(' 1', 9)

(' 6', 2)

(' 34', 3)

(' 4', 3)

(' 5', 13)

(' 8', 6)

(' 53', 18)

(' 19', 2)

(' 38', 4)

(' 51', 2)

(' 31', 4)

(' 18', 3)

(' 76', 5)

(' 77', 4)

(' 74', 1)

(' 75', 6)

(' 72', 3)

(' 73', 2)

(' 70', 42)

(' 71', 3)

(' 90', 9)

(' 100', 2)

(' 78', 1)

(' 7', 3)

sum = 1128