Hi Guys,
Sorry for delay, I was attempting to write-up some of the statistical methods and processes, that allow me to add confidences scores to similar campaign and propensity scores to a profiles propensity to engage with a campaign.

Here's a quick overview and example and I'll write-up more detail explanations by documenting the process of developing a cluster, scorecard, profile, and model.

So we want to cluster by ip address and pull as many distinct characteristics from that ip as possible.
This will allow us to compare that ip to other ips and create useful clusters.

Let's look at some example objects and processes based off our current data set and ad-ops procedures (liverRail data).
Then we'll identify any limitations, blockers, missing data and how we can resolve those issues.

Profiling
Score Card development
Clustering
Campaign modeling

Profiling
The process of resolving as many distinct characteristics from an ip (cell phone: dpidsha1, GPS). I'll send in a follow-up email all the available properties that I know how to profile on, based on my work at http://voltari.com and http://ociweb.com/ (I also had to attend OCI certification classes for their ACE and TAO frameworks). As I mentioned before, Their products are past beta and currently running on openx and nexxage RTB exchanges and also if I choose not to continue I will have references from the CTO's as well as human resources. The processes I am building for Impaktu are at production state at other companies and returning 70% or greater confidence of engagment.

Profile Object
Currently based off current data collection results)
Each data point will need to pass through a weighting formula, which among other things includes the number of other times this profile has engaged with that data point.
For example the vertical 'automotive' might be added as a runtime parameter the vast tag, along with a number of other verticals.
My tracker will pick up that vertical during engament(impression,click—through,etc).
But how much did that particular vertical, category, etc. influence the profile decision to engage.
Well we have a series of formulas that result in a weight for that datapoint in the profile:

Name	Descr
Convergence Tolerance	Value considered acceptable for convergence
Maximum Iterations for Convergence	Maximum number of iterations to perform to achieve convergence

Number of Constants	Number of constant terms to employ in the model
Autocorrelation	Boolean flag denoting whether to employ autocorrelation correction to the model
Rho Tolerance	NULL
Rho Significance	NULL
Maximum Rho Iterations	NULL
Functional Form	Functional form to apply to the model. Can be LINEAR, LOG, LOGRHS, or LOGLHS
R Squared Uncorrected	NULL
R Squared Corrected	NULL
Durbin Watson	NULL
Sum Squares Total (SST)	NULL
Sum Squares Regression (SSR)	NULL
Sum Squares Error (SSE)	NULL
Standard Error (SE)	NULL
Chi Square Coefficient	NULL
Initial Log Likelihood Function	NULL
Log Likelihood Function	NULL
Prediction Success Index	NULL
Corrected L Ratio Index	NULL
Uncorrected L Ratio Index	NULL
Number Residuals GT 0.5	NULL
Number Iterations for Convergence	NULL
Maximum Attempts Per Iteration	NULL
Number Observations	Number of Observations
Number Missing	Number of Missing Observations
Min Value	Min Value of Dependent Variable
Max Value	Max Value of Dependent Variable
Mean	Mean of Dependent Variable
Median	Median of Dependent Variable
Variance	Variance of Dependent Variable
Std. Deviation	Standard Deviation of Dependent Variable
F-statistic	Overall Model Significance F-statistic
Number of Independent Variables	Number of Independent Variables
Percent Holdout	Percent of the modeling sample to hold out
Min Reference Size	Minimum size of the reference population
Max Reference Size	Maximum size of the reference population
Reference Factor	Factor to apply to target popluation to determine reference population

These processes are also used for modeling and propensity determination.

You can see some things in the list that make sense without a statistical background:

- Standard deviation
- Max/min referernce size
- Number of independent variables
- Number of observations
- Likelihood functions
- Prediction success index

There a free open source called 'R' which can help you visualize this statistical analysis. I'm not sure when we'll get to that though. (one-man teaming it here ☺).

So let's look what a current a profile will be analyzed using, what are the limitations and how we can add to this dataset.

1) **UserAgent**
    - Browser (/w version)
    - Operating System (/w version)
    - Table or Iphone (/w version)
2) **Ad Position**
    - Could be useful if it shows a statistically significant re-occcurence.
    - Combine that with view % and maybe the user will watch 100% if it's post-roll
3) **Tags**
    - Keywords added by AdOps to describe the campaign
    - As I explained earlier, these could also be useful if we can use certain methods to isolate the re-occurrence of the individual words across campaigns and add weights to them. So a very oversimplification of the process would be that the tag 'hair care' is added to multiple campaigns, in the presence of varying other keywords. We'll run several stat methods to see if the keyword 'hair care' should have a weight to it. FYI, there are multiple steps in determining this, another would be to compare this profile to others in the cluster and determine if removing the keyword makes a statistical difference. So we start using attributes like number observations, number of independent variables and running these properties through multiple likelihood, distance (similarity), and predictor stat methods. This should probably be in the 'process' section but it's a first-impression thing, I don't want you to think I just assign hair care to the profile. ☺
4) **Verticals**
    - Keywords added by AdOps(publisher) to describe page contents.
    - Same methodology applies here as with Tags
    - If we can statistically isolate the fact that a profile visits 'auto racing' web pages, we can assign the characteristics of an auto racing fan to the profile. The 'likelihood' that it's a male age x-y, etc. This is done in correlation with additional profile info. In later sections or follow up documentation we will discuss 'Limitations' and how to get around them. Gathering the actual age of the profile will be one of those considerations. This is one generalized approach that could help, but we can get more specific and closer to that actual info.
5) **Categories**

- IAB domain level categories
- See Tags and Verticals

6) **Content**
   - Number describing content of the environment the video is playing in.
   - Seems very generic but the processing will determine If it's a statistical factor and give it a weight (may get a weight of 0.0001 or something).
7) **URL**
   - Url of page embedding the video
   - Should probably develop or integrate our own site characteristics db
   - Maybe using cookies showing how this page is accessed and it's characteristics
   - Again those characteristics weighted across number of observations, etc can be assigned to profile.

Results

Profile

{
Ip address
-User Agent
Ad Position
Tags
Verticals
Categories
Content
URL
}

Limitations

What would increase accuracy would be data points like:

PredictorTypeId	Descr
0	NotPredictor
21	Psychographics
22	Income
23	HH Characteristics
24	Lifestyle
25	Charitable Contributor
26	Mobility
27	Age
28	Person Characteristics
29	Multicultural
35	Education
36	Gender
37	Marital Status

38	Credit
39	Home Value
41	Device OS
62	Travel
75	Segment Group
76	Segment Sub-group
77	Occupation
78	Second Party
79	Custom
80	Household Size
81	Contributions
82	Interests
83	Reading
84	Travel
85	Pets
86	Parenting
87	Spectator Sports
88	Collectibles
89	Hobbies
90	Home
91	Net Worth
92	Dwelling
93	Household
94	Personix Hispanic
95	Vehicle
96	Social Influence
97	Mobile Social Networker
98	Facebook
99	Twitter
100	Linked In
101	You Tube
102	Poster
103	Video
104	Race/Ethnicity
105	Language
106	Own/Rent
107	Length of Residence
108	Occupation
109	Personicx
110	Personicx Digital
111	Personicx Digital (Groups)

- You’ll see some of the predictor numbers skipped as I deleted proprietary categories from a previous table.

Solutions

Well need to develop some methods for obtaining more profile(user/cluster) specific data.

- Age,race,occupation,etc.

This could involve:

1) cookies
2) additional javascript calls
3) deeper publisher integeration
- requiring certain macros are available on page (LR_TITILE)
- configuring players to asynchronously send additional data
- Overriding macros with user info (login,usernames,etc), LiveRail will then send that info.
- I think I sent a comprehensive list earlier this year, update and add that to a follow up email.
4) The best solution, I’ve seen over 3 companies personally (qualityhealth.com, IAC, Voltari) is to wrap the adserver tag in a proprietary Impaktu tag. This allows us to :
- Receive first party data like the referrer
- Add additional javascript to the adtag to collect all available information.
- I haven’t reviewed all the intricacies of replacing a VAST tag but I personally travelled to many of the IAC properties like Match.com,Evite and CitySearch and retagged their sites(each running different code bases such as c#, Java, C++ etc, as well as remotely accessing many other properties to do the same. At the time of switch over to IAC servers, there was zero impression lost. The President at that time eventually became CEO of Voltari and ask me to move to that company. So althought I don’t have actual VAST experience the methodologies are the same, and I’m 100% confident I could setup a Impaktu ad tagging system.

Score Cards

The next things we’ll need to review are score cards.
Everything gets a score card. And a score card is really just a set of values and associated weights.
The score card allows me to statistically group ip addresses into clusters.