

SAP Big Data Analytics on Mobile Usage

A George Mason University study

Students: Arturo Buzzalino, Tanner Suttles, Mitul Patel, Justin Nguyen

Sponsor: Steve Garcia

May 9, 2014

Public

The SAP logo is located in the bottom left corner. It consists of the letters "SAP" in a bold, white, sans-serif font, set against a blue trapezoidal background that tapers to the right.The George Mason University logo is located in the bottom right corner. It features a stylized green leaf icon above the text "GEORGE MASON UNIVERSITY" in a green, serif font. The text is arranged in three lines: "GEORGE" on the top line, "MASON" in a larger font on the middle line, and "UNIVERSITY" on the bottom line.

Overview

- Problem Definition and Scope
- Data Overview
- Reduced Scope
- Data Analysis
- Modeling Techniques
- Data Processing for Model
- Model Results
- Conclusions
- Further Research



Background - SAP Consumer Insight 365

SAP is developing a new product to enable businesses to better manage and expand their markets

Mobile carriers have an enormous amount of unused consumer data

- When leaving your home you take your keys, wallet and **phone**
- Mobile devices are constantly producing data outlining a consumer's lifestyle

This data can be key for any business to boost growth

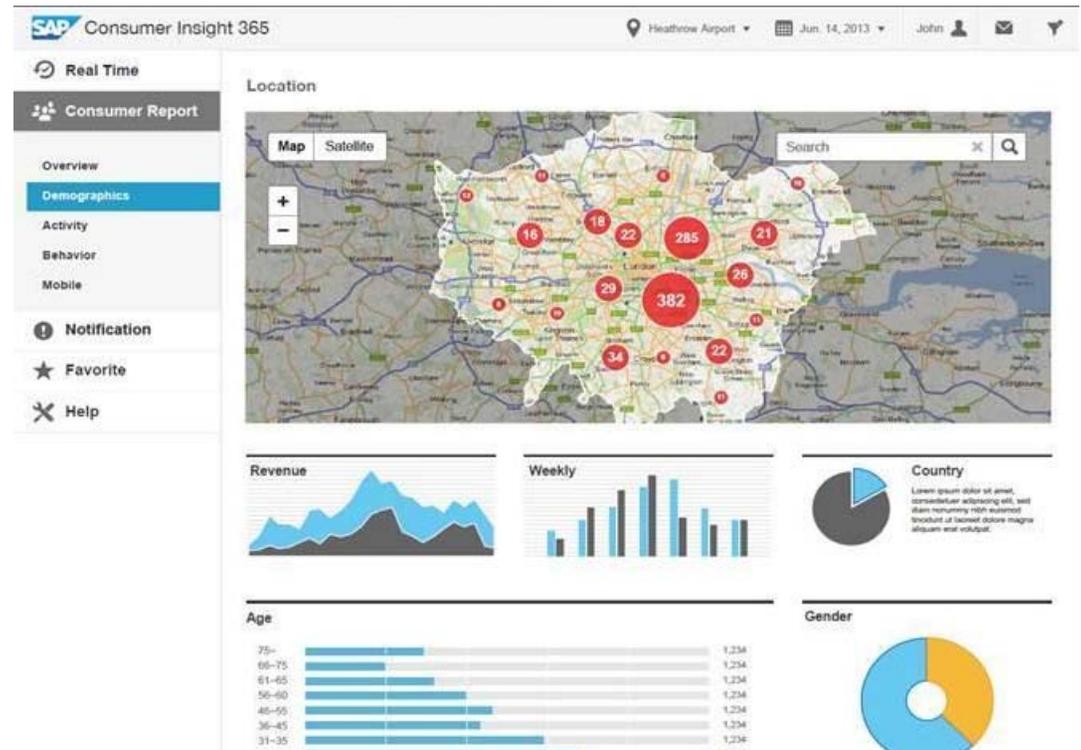
- Focalize marketing efforts
- Who are your potential customers?
- Thousands of applications
 - Growing App market, traffic patterns, malls, airports etc.

Background - SAP Consumer Insight 365

Mobile Carriers can monetize data, as well as gain insights on their own consumer base

Consumer Insight 365 is a tool able to put this data to work

- Texting, calling habits
- Geo-location and socio-demographics
- Malls, airports, attractions footfall
 - Who is frequenting? For how long?
- Stores
- Interests: Facebook, Twitter, URL categories



GMU Project - Problem Definition

SAP receives anonymized data

- Age and gender of the plan holder, will be provided by the carrier
- This information is not always known
 - Especially unknown when it is a prepaid plan

The team's role will be to identify patterns that suggest age and gender of user

- Texting / calling habits; Geolocation; Point of Interest (POI e.g. Starbucks); URL categories
- Possible consideration:
 - Socio-demographic based on location; time spent in POI;

Scope

Objective:

The team will utilize data provided by the mobile carrier to determine usage pattern from the population of known age group & gender to predict the unknown population of ~~age group & gender~~. ~~The data to be analyzed spans one month, 1.4 million users, and over 1 billion rows of data.~~

Deliverables:

Data Model

- Imply ~~age and~~ gender of the user, important for marketing
- Input includes ~~Texting / calling habits; Geolocation; Point of Interest (POI e.g. Starbucks); URL categories~~

Report

- Description of pattern that lead to model
- Description of Model
- Sensitivity Analysis
- Report on inferred ~~age and~~ gender of mobile users

Summary of Relevant Data

Received data late – did not get data until March 7, 2014

Received:

- 5 Days of data
- User Information
 - Age band
 - Gender
 - Home zip code
 - Handset
- Web Activity
 - URL Domain
 - Start/end times visiting each domain
 - Bytes Transferred

Missing:

- Call
- Text

Data Overview

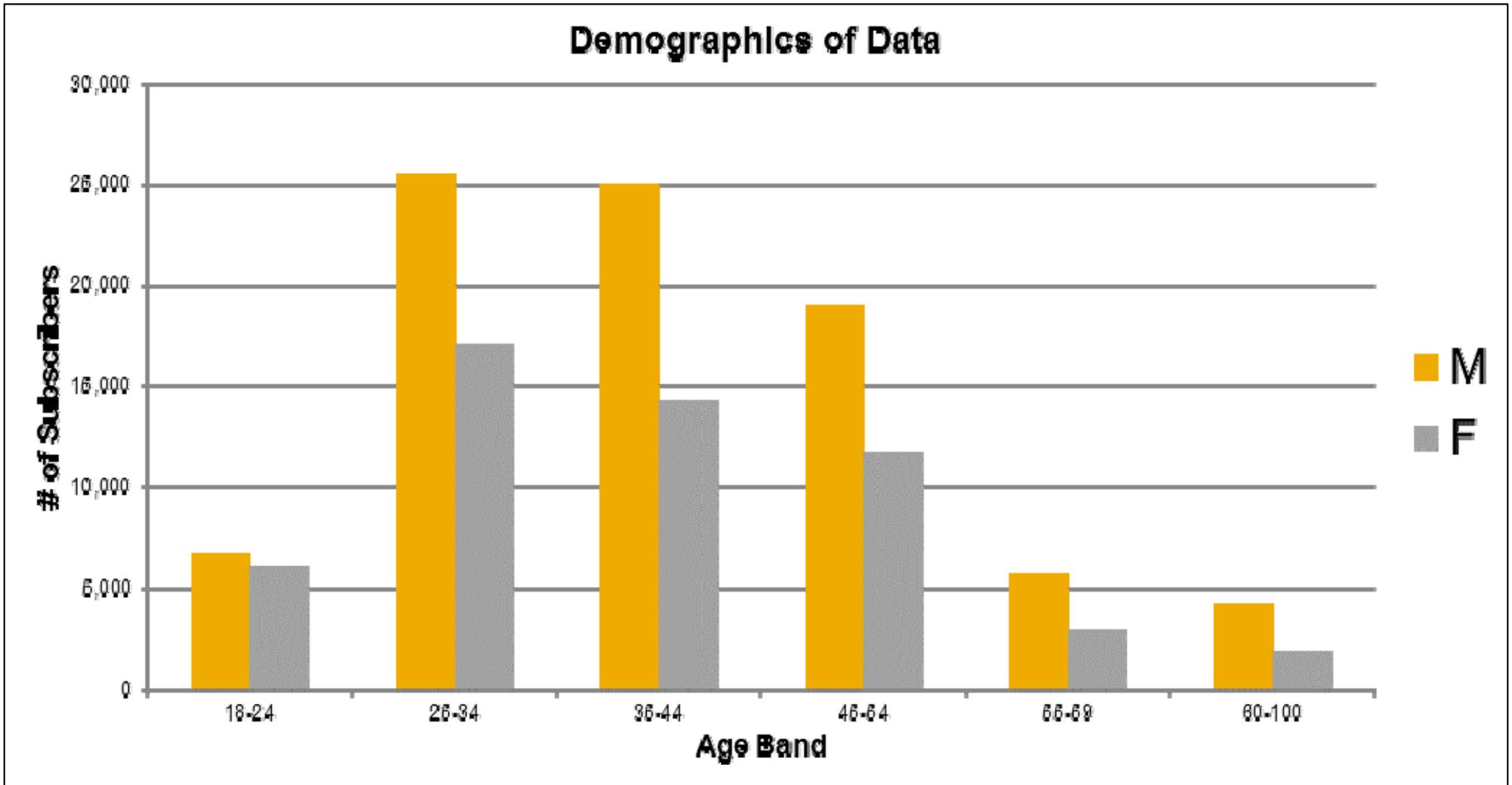
Production

3 distinct data sets:

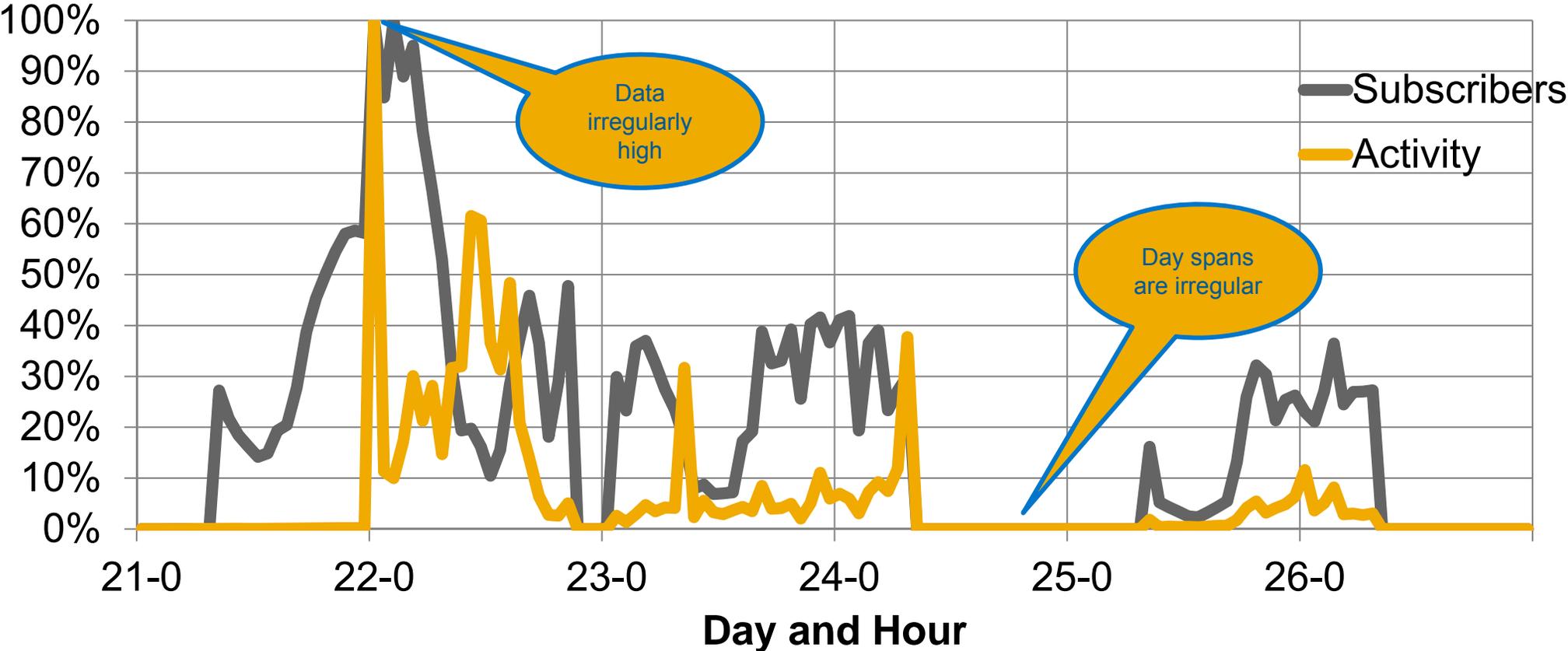
- Learning/training dataset – Used for our algorithms to train algorithms
- Testing/verification data set – Male & Female is known, algorithm will be tested
- Unknown gender data – The model will be tested and confidence will be provided

Males	86,334
Females	54,075
Unknowns	63,444

Data Distribution



Users vs Data



Handset and Domain Categorization

- A domain categorization API (3rd party company) was used to assign categories
- Re-categorized two of the larger generic categories

Technology - Other		
Percent	DOMAIN	NEW CATEGORY
62%	'meta.radioactive.sg'	Radio
3%	'ping.chartbeat.net'	Marketing Services
3%	'data.gosquared.com'	Marketing Services
2%	'www.azonano.com'	News
2%	'armdl.adobe.com'	App Updater
2%	'up.cm.ksmobile.com'	App Updater
1%	'cs.atdmt.com'	Online Ads - Other
1%	'www.instapaper.com'	Offline Website
1%	'mobilizer.instapaper.com'	Offline Website
1%	'ads.radioactive.sg'	Radio
1%	'apps.radioactive.sg'	Radio
1%	'oc.umeng.com'	Marketing Services

- A large number of the handsets were irregular and too specific
 - The team broke these into more encompassing sets
 - i.e. Sony, Apple, Samsung without specifics

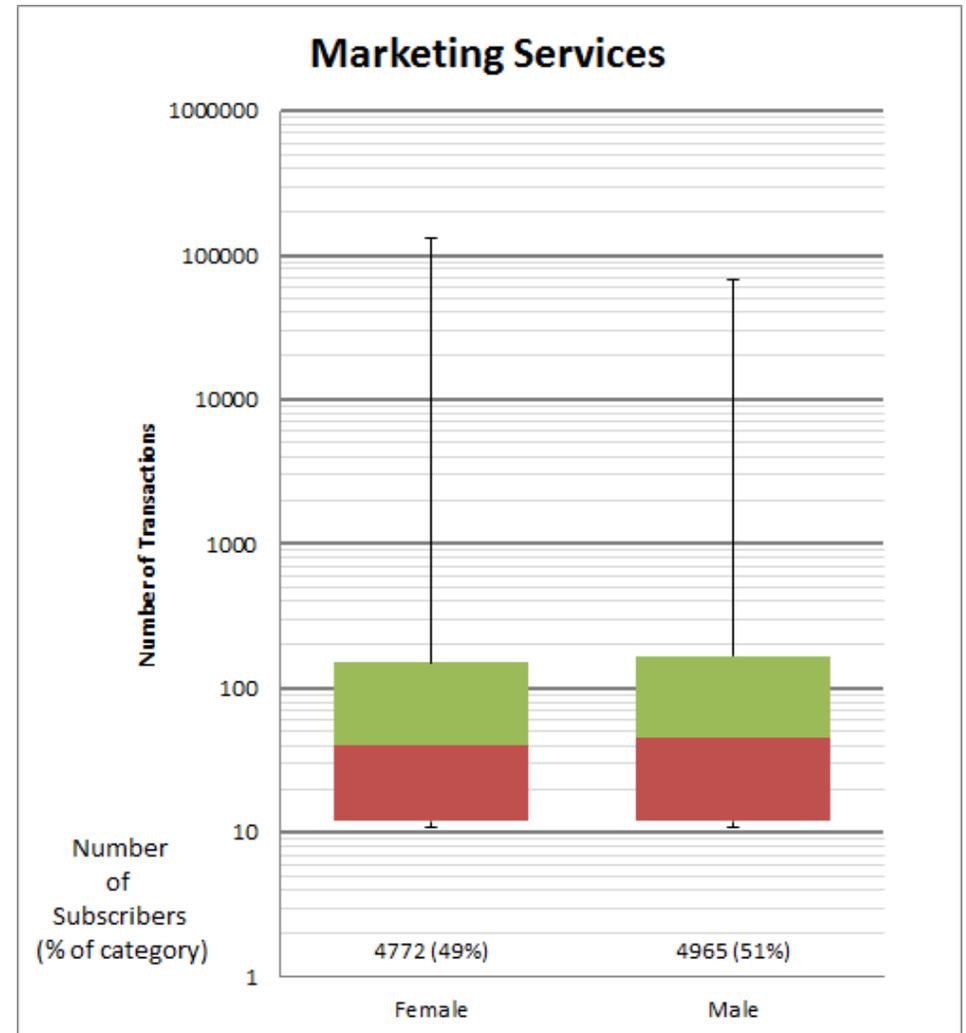
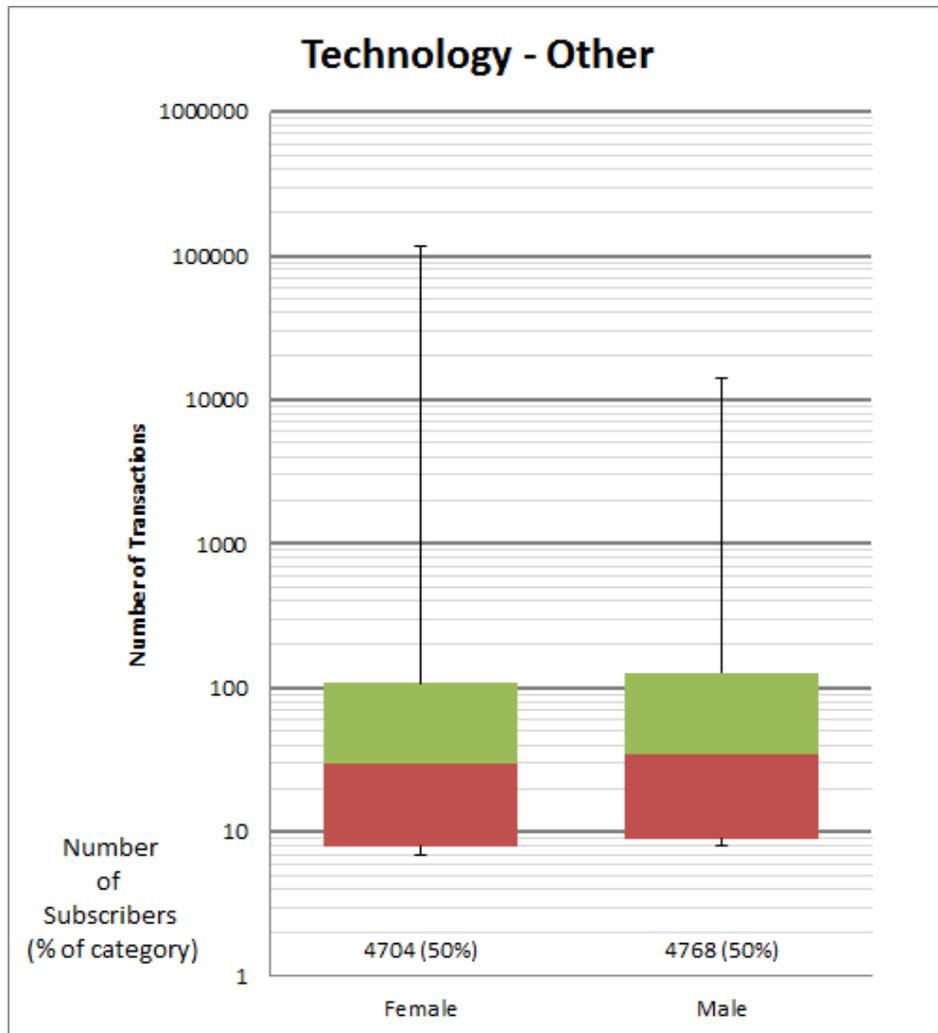
Top Visited Categories

- Visitors drop off quickly
- Top categories don't differentiate

Rank	Category	% of Users Visited	Female / Male	
1	uncategorized	68%		0%
2	Online Ads - Other	50%		0%
3	Marketing Services	49%		1%
4	Technology - Other	47%		0%
5	Content Server	37%		1%
6	Games	21%		-3%
7	News	16%		6%
8	Portal Sites	16%		-3%
9	Information Security	16%		4%
10	File Repositories	16%		-1%
11	Streaming & Downloadable Video	15%		-1%
12	Business - Other	15%		1%
13	Computer Peripherals	14%		2%
14	Personal Pages & Blogs	12%		-1%
15	Entertainment - Other	11%		0%
16	Travel - Other	11%		1%
17	Community Forums	11%		1%
18	Social Networking	11%		1%
19	Mobile Phones	10%		1%
20	Online Shopping	9%		-3%

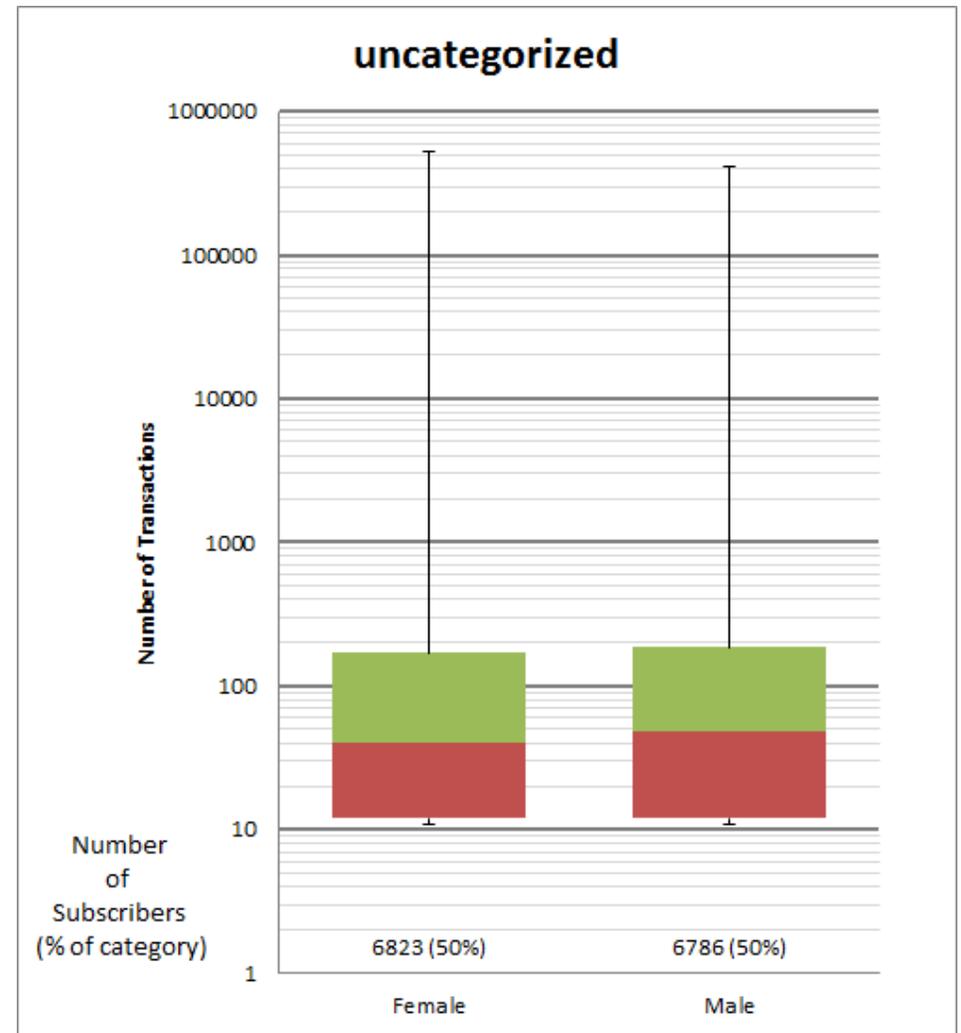
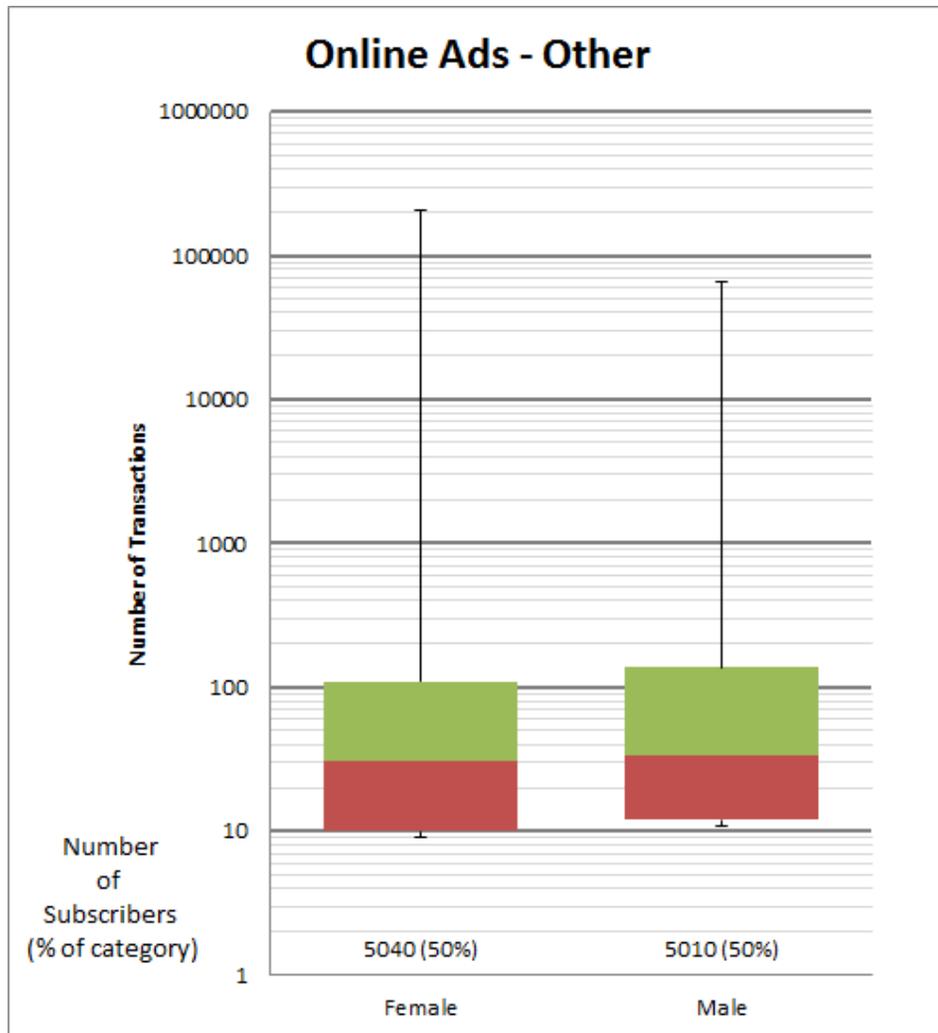
Top Visited Categories [2]

- Number of transactions for many categories were very similar for each gender



Top Visited Categories [3]

- Number of transactions for many categories were very similar for each gender



Top Gender Difference Categories

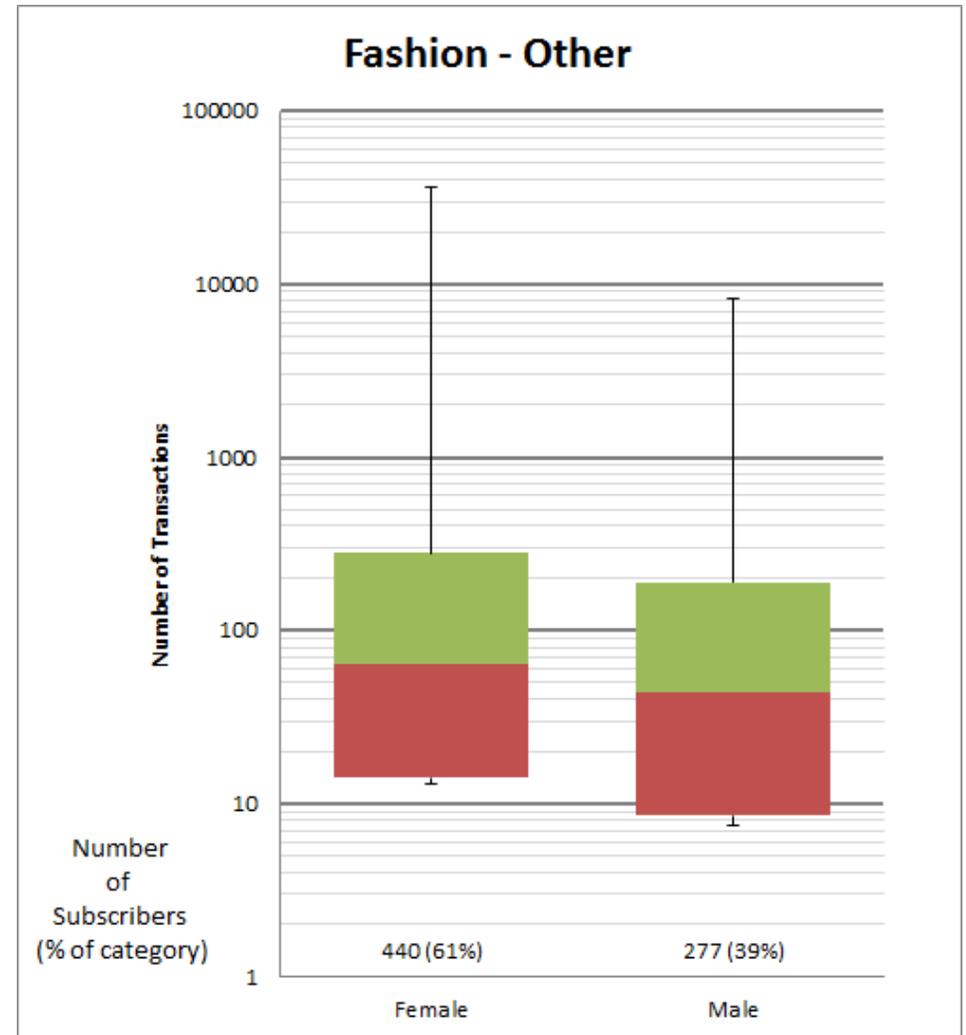
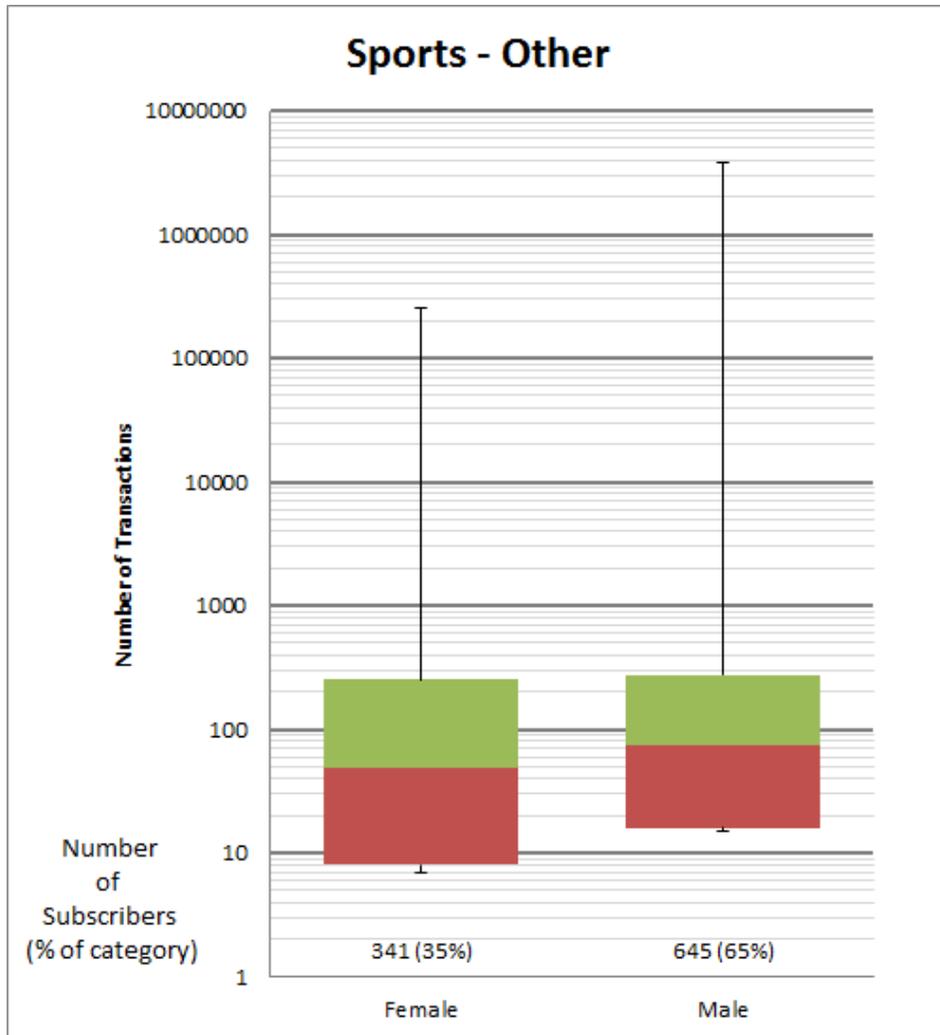
- Categories with gender bias had relatively few visitors (<1%)

Rank	Category	% of Users Visited	Female / Male
25	Pornography	6%	14%
27	Unreachable	5%	9%
29	Sports - Other	5%	15%
33	Fashion - Other	4%	-11%
41	Gambling	3%	11%
48	Radio	2%	-17%
53	Arts - Other	2%	-10%
54	Dating & Relationships	2%	13%
58	Malware Distribution Point	1%	12%
59	Educational Institutions	1%	-10%
59	R-Rated	1%	11%
61	Cartoons & Anime	1%	9%
62	Instant Messenger	1%	15%
64	Piracy & Copyright Theft	1%	17%
67	Sex & Erotic	1%	25%
68	Construction	1%	-9%
73	Legal Issues	1%	11%
74	Product Reviews & Price Comparisons	1%	16%
76	Gay	1%	28%
78	Home & Garden - Other	1%	-15%

*Categories with greater than 8% swing in gender

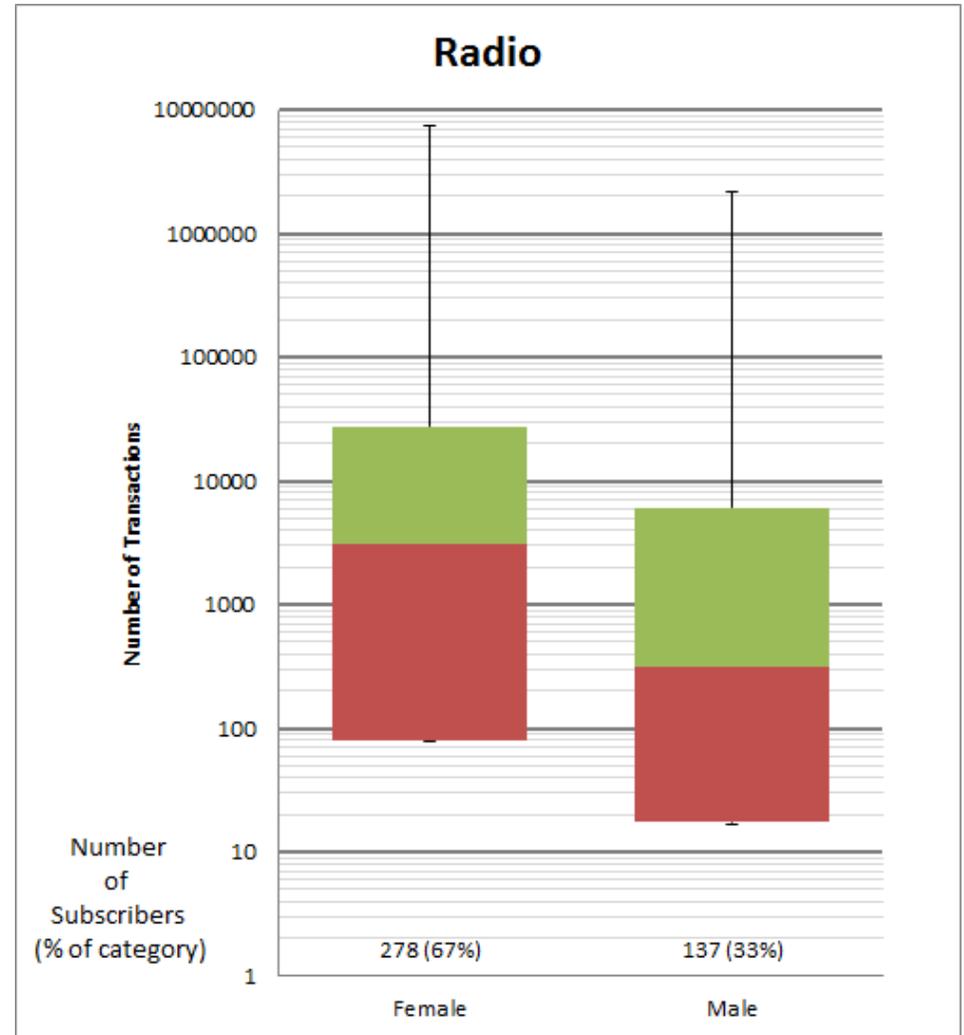
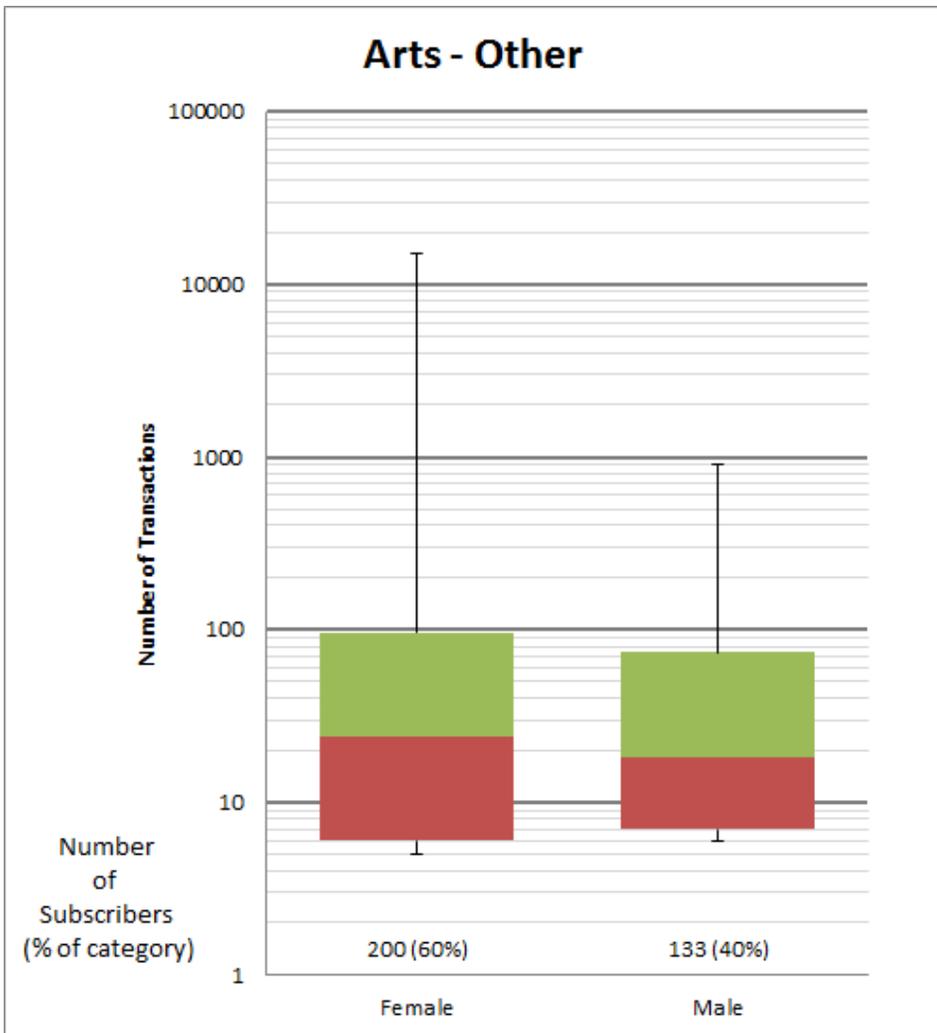
Top Gender Difference Categories [2]

- Number of transactions for many categories were very similar for each gender



Top Gender Difference Categories [3]

- Number of transactions for many categories were very similar for each gender



Classification Algorithms

- **Naive Bayes**

- Doesn't consider relationships between attributes
- Based on conditional probabilities.
- Finds the probability of an event occurring given the probability of another event that has already occurred.

- **Chi-squared Automatic Interaction Detector (CHAID)**

- Constructs non-binary trees for classification problems
- Relies on the Chi-squared test to determine the best next split at each step
- If the test shows a pair of predictors is not statistically significant, the predictor categories will merge

Data Processing

- Correlation Analysis
 - Box plots
 - Bar graphs
 - Choosing parameters to evaluate
- Select Subscribers – **SAP HANA Studio**
 - Remove nulls (domain and handset)
 - Look for anomalies and unusual activities (ex. Radio)
- Model Data – **SAP HANA Studio**
 - Count transactions for each person
 - 5 minute interval durations
 - Binary decision for categories
- Pivot – **SAP Data Services Designer**
 - Pivoting URL categories to have one gender result per user
- Model Development – **SAP HANA PAL**
 - Naive Bayes and CHAID Algorithms
 - Training Set for Algorithm Learning (class attribute)
 - Testing Set for Algorithm Evaluation (assigning # IDs)
- Model Evaluation
 - Sorting IDs and mapping back to dataset for accuracy
 - No confidence provided by HANA PAL

Phase 1: Raw Transactions

- Data Set Definitions
 - Training Set 1 = 10,000 distinct subscribers for each gender w/ random age bands (20k total)
 - Testing Set 1 = 500 distinct subscribers for each gender w/ random age bands (1k total)
- Evaluates category activity counts
- Results would have different gender results for same subscribers due to multiple rows
- Need to manipulate training and testing sets to have one result per subscriber

SUB_ID	CATEGORY	TRANSACTIONS	GENDER
1	Sports	3487	M
1	Gambling	34	M
1	News	4356	M
2	Shopping	23	F
2	News	123	F
3	Technology	7658	M
3	Games	154	M

- No results due to multiple predictions for one subscriber

Phase 2: Pivoted Category with Activities and Duration Span

- Data Set Definitions
 - Training Set 1 = 10,000 distinct subscribers for each gender w/ random age bands (20k total)
 - Testing Set 1 = 500 distinct subscribers for each gender w/ random age bands (1k total)
- Pivoted 150 categories w/ activity and duration span values
- Also evaluates home zip and handset

Train Set	Test Set	Algorithm	Total Accuracy
1	1	Bayes	50%
1	1	CHAID	55%

- Bayes inferred all females
- HANA PAL does not provide a confidence of results
- Too many correlation parameters
- Need to go with simpler approach

SUB_ID	HOME_ZIP	HANDSET	ART_ACTIVITIES	ART_DURATION_SPAN	...	GENDER
1	710	Sony	0	0	...	M
2	540	Samsung	1763	15	...	F
3	679	Apple	0	0	...	M

Phase 3: Pivoted Category with Binary Activities

- Data Set Definitions

- Training Set 1 = 10,000 distinct subscribers for each gender w/ random age bands (20k total)
- Testing Set 1 = 500 distinct subscribers for each gender w/ random age bands (1k total)
- Training Set 2 = 20% of all distinct age band of all subscribers for each gender (14k M, 14k F, and 28k total).
- Testing Set 2 = 3500 distinct subscribers for each gender w/ random age bands (7k total)

- Categories

- 1 VISITED
- 0 NOT VISITED

- Home zip and handset in training sets made results worse

Train Set	Test Set	Algorithm	Total Accuracy
1	1	Bayes	62%
1	1	CHAID	50%
2	2	Bayes	55%
2	2	CHAID	50%

- Bayes not inferring all females anymore
- CHAID is better with continuous parameters
- Binary approach works well with Bayes

SUB_ID	ART_ACTIVITIES	NEWS_ACTIVITES	...	GENDER
1	0	1	...	M
2	1	1	...	F
3	0	0	...	M

Phase 4: Master Test Set for Three Training Sets

- Train 1
 - Selected 10,000 distinct subscribers for each gender
 - No regard for age band
- Train 2
 - Selected approximately 20% of data per age band
 - 50/50 genders
- Train 3
 - Selected 1000 subscribers from each age band and gender
- Master Test Set
 - Selected 500 distinct subscribers for each age band and gender that are not in the other three training sets

Train Set	Test Set	Algorithm	Male Accuracy	Female Accuracy	Total Accuracy
1	Master	Bayes	56%	53%	54%
1	Master	CHAID	52%	52%	52%
2	Master	Bayes	57%	54%	55%
2	Master	CHAID	52%	51%	52%
3	Master	Bayes	58%	54%	55%
3	Master	CHAID	51%	52%	52%

- Bayes accuracy of male predictions were higher because it inferred females more often

- Results are all too similar even with different demographic training data

Phase 5: Subsampling Subscribers

- Data Set Definitions

- Training Set 2 = 10,000 total distinct subscribers with only top 27 gender differentiating categories
 - Removal of users outside of categories reduced the total from 28k to 10k
- Testing Set 1 = 1000 total distinct subscribers with only top 27 gender differentiating categories
 - Same amount of subscribers from original set
- Master Testing Set = 3,200 total distinct subscribers with only top 27 gender differentiating categories
 - Removal of subscribers outside of categories reduced the total from 6k to 3.2k

Train Set	Test Set	Algorithm	Total Accuracy
2	1	Bayes	62%
2	1	CHAID	50%
2	Master	Bayes	55%
2	Master	CHAID	52%

- Same results when narrowing top categories
- Same results from full category models
- Testing sets have more impact on results
- Demographics in training set did not seem to impact results

Conclusions

- Main delivery change is a model capable of inferring only gender and the exclusion of an age inferring algorithm
 - Roadblocks and data issues
 - Bad data
- The delivery includes details of generating the model, its accuracy, and its sensitivity under varying training and testing scenarios
- Data integrity was of high interest to SAP
 - The team participated in exposing and summarizing these findings
 - Turned out to be an unforeseen deliverable

Further Research

- The immediate next step is for SAP to run the model delivered on the new 1 month data
 - This new set is complete, and does not have the data issues the team faced
 - The addition of call, text, and location data
- Run categorization API on the full URN path of the websites, instead of just domain
 - Higher granularity to expose the differences in the genders
 - `cnn.com/basketball` vs `cnn.com/finance`
- Attempt furthering analysis with data analysis experts within SAP and using other software
 - R, SPSS, and SAP 3rd party partners



Thank you
Question?

© 2014 SAP AG or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP AG or an SAP affiliate company.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP AG (or an SAP affiliate company) in Germany and other countries. Please see <http://global12.sap.com/corporate-en/legal/copyright/index.epx> for additional trademark information and notices.

Some software products marketed by SAP AG and its distributors contain proprietary software components of other software vendors.

National product specifications may vary.

These materials are provided by SAP AG or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP AG or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP AG or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP AG or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP AG's or its affiliated companies' strategy and possible future developments, products, and/or platform directions and functionality are all subject to change and may be changed by SAP AG or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, which speak only as of their dates, and they should not be relied upon in making purchasing decisions.

© 2014 SAP AG oder ein SAP-Konzernunternehmen. Alle Rechte vorbehalten.

Weitergabe und Vervielfältigung dieser Publikation oder von Teilen daraus sind, zu welchem Zweck und in welcher Form auch immer, ohne die ausdrückliche schriftliche Genehmigung durch SAP AG oder ein SAP-Konzernunternehmen nicht gestattet.

SAP und andere in diesem Dokument erwähnte Produkte und Dienstleistungen von SAP sowie die dazugehörigen Logos sind Marken oder eingetragene Marken der SAP AG (oder von einem SAP-Konzernunternehmen) in Deutschland und verschiedenen anderen Ländern weltweit. Weitere Hinweise und Informationen zum Markenrecht finden Sie unter <http://global.sap.com/corporate-de/legal/copyright/index.epx>.

Die von SAP AG oder deren Vertriebsfirmen angebotenen Softwareprodukte können Softwarekomponenten auch anderer Softwarehersteller enthalten.

Produkte können länderspezifische Unterschiede aufweisen.

Die vorliegenden Unterlagen werden von der SAP AG oder einem SAP-Konzernunternehmen bereitgestellt und dienen ausschließlich zu Informationszwecken. Die SAP AG oder ihre Konzernunternehmen übernehmen keinerlei Haftung oder Gewährleistung für Fehler oder Unvollständigkeiten in dieser Publikation. Die SAP AG oder ein SAP-Konzernunternehmen steht lediglich für Produkte und Dienstleistungen nach der Maßgabe ein, die in der Vereinbarung über die jeweiligen Produkte und Dienstleistungen ausdrücklich geregelt ist. Keine der hierin enthaltenen Informationen ist als zusätzliche Garantie zu interpretieren.

Insbesondere sind die SAP AG oder ihre Konzernunternehmen in keiner Weise verpflichtet, in dieser Publikation oder einer zugehörigen Präsentation dargestellte Geschäftsabläufe zu verfolgen oder hierin wiedergegebene Funktionen zu entwickeln oder zu veröffentlichen. Diese Publikation oder eine zugehörige Präsentation, die Strategie und etwaige künftige Entwicklungen, Produkte und/oder Plattformen der SAP AG oder ihrer Konzernunternehmen können von der SAP AG oder ihren Konzernunternehmen jederzeit und ohne Angabe von Gründen unangekündigt geändert werden.

Die in dieser Publikation enthaltenen Informationen stellen keine Zusage, kein Versprechen und keine rechtliche Verpflichtung zur Lieferung von Material, Code oder Funktionen dar. Sämtliche vorausschauenden Aussagen unterliegen unterschiedlichen Risiken und Unsicherheiten, durch die die tatsächlichen Ergebnisse von den Erwartungen abweichen können. Die vorausschauenden Aussagen geben die Sicht zu dem Zeitpunkt wieder, zu dem sie getätigt wurden. Dem Leser wird empfohlen, diesen Aussagen kein übertriebenes Vertrauen zu schenken und sich bei Kaufentscheidungen nicht auf sie zu stützen.