Click it!

Team 3 Trees

Chung Lau, Anita Miller, Steve Nelson, Todd Schultz, Paul Scarpa,
Srinivasa Ramasamy, Naresh Jarugala

Avazu CTR Kaggle competition

- **Predict whether a mobile ad will be clicked**
- **Scored using a log loss metric**
- **Provided 10 days worth of data for training**
- **Provided 1 days worth of data for testing**

Log loss scoring

$$LogLoss = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)]$$

- Brutal for extreme wrong predictions
- Defines critical performance indicator
- Can the score be increased without accuracy improvements?

Best Score = 0.3990766

- Platform = Azure ML
- Algorithm = Boosted decision tree
- Maximum leaves per tree = 20
- Minimum samples per leaf = 10
- Learning rate = 0.2
- Number of trees = 100
- Entire training set

- Random guessing benchmark score = 0.6931472
- https://www.kaggle.com/c/avazu-ctr-prediction/leaderboard

Divide, conquer, and ensemble!

- Work separately but share results as you go
- Each create our own unique models
- Ensemble together at the end with a simple average of the probabilities

- Strategy determined by demonstrated power of ensembles and the reality of the logistics of organizing the team

train

500,048 observations of 24 variables

	id	click	hour	C1	banner_pos	site_id	site_domain	site_category	app_id	app_domain	app_category	device_id	device_ip	device_model	device_type	device_conn_type	C14	C15	C16	C17	C18	C19	C20	C21
1	1.000009e+18	0	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	ddd2926e	44956a24	1	2	15706	320	50	1722	0	35	-1	79
2	5.550643e+18	1	14102100	1002	0	763a42b5	4c26e9ba	50e219e0	ecad2386	7801e8d9	07d7df22	60ac4d27	c52fa209	81a9e2c3	0	0	15908	320	50	1752	3	297	100081	82
3	1.804285e+19	1	14102100	1005	0	543a539e	c7ca3108	3e814130	ecad2386	7801e8d9	07d7df22	a99f214a	e31dd869	dab059c4	1	0	20362	320	50	2333	0	39	-1	157
4	3.803035e+18	1	14102100	1005	0	d9750ee7	98572c79	f028772b	ecad2386	7801e8d9	07d7df22	a99f214a	32d5f89e	684581ce	1	0	18094	320	50	2060	3	39	-1	23
5	1.480825e+19	1	14102100	1005	0	85f751fd	c4e18dd6	50e219e0	e2fcccd2	5c5a694b	0f2161f8	a99f214a	b6d9dc16	6332421a	1	0	6616	320	50	576	2	35	-1	32
6	3.095319e+17	1	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	086ef670	2203a096	1	0	15706	320	50	1722	0	35	-1	79
7	1.101398e+19	1	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	78090c13	8a4875bd	1	0	15703	320	50	1722	0	35	-1	79
8	4.236099e+18	1	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	82702678	2203a096	1	0	15708	320	50	1722	0	35	100084	79
9	1.114372e+19	1	14102100	1005	0	85f751fd	c4e18dd6	50e219e0	e2fcccd2	5c5a694b	0f2161f8	a99f214a	313b3224	1ccc7835	1	0	20633	320	50	2374	3	39	-1	23
10	7.070886e+18	1	14102100	1005	1	93de26ae	7d05db75	335d28a8	ecad2386	7801e8d9	07d7df22	a99f214a	211a11af	a4b0a5ea	1	0	20596	320	50	2161	0	35	-1	157
11	1.238268e+19	1	14102100	1005	1	fe8cc448	9166c161	0569f928	ecad2386	7801e8d9	07d7df22	a99f214a	40e6c535	27f0942f	1	0	18993	320	50	2161	0	35	-1	157
12	3.860264e+18	1	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	ae1f065e	2203a096	1	0	15708	320	50	1722	0	35	100084	79
13	5.899984e+18	1	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	75bb1b58	23823d8b	1	0	15704	320	50	1722	0	35	-1	79
14	2.701318e+18	1	14102100	1005	1	6bdcda77	b0a0505f	72722551	ecad2386	7801e8d9	07d7df22	a99f214a	32c9e9ee	900981af	1	2	20596	320	50	2161	0	35	100034	157
15	1.625072e+19	1	14102100	1010	1	85f751fd	c4e18dd6	50e219e0	a607e6a7	7801e8d9	0f2161f8	d979cf01	ff0393c6	f07e20f8	4	0	21665	320	50	2493	3	35	-1	117
16	5.147084e+18	1	14102100	1005	0	da79c793	71ed77a0	f028772b	ecad2386	7801e8d9	07d7df22	a99f214a	4aaba82e	84ebbcd4	1	0	18993	320	50	2161	0	35	100210	157
17	4.978811e+17	1	14102100	1005	0	543a539e	c7ca3108	3e814130	ecad2386	7801e8d9	07d7df22	a99f214a	28444dcf	24f6b932	1	0	20352	320	50	2333	0	39	-1	157
18	1.570463e+19	1	14102100	1005	1	e151e245	7e091613	f028772b	ecad2386	7801e8d9	07d7df22	a99f214a	88f84b1f	8a4875bd	1	0	20362	320	50	2333	0	39	-1	157
19	4.459188e+18	1	14102100	1005	1	5ee41ff2	17d996e6	f028772b	ecad2386	7801e8d9	07d7df22	a99f214a	7232a832	88fe1d5d	1	0	19771	320	50	2227	0	687	100077	48
20	7.177161e+18	1	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	97d00d2a	8a4875bd	1	0	15705	320	50	1722	0	35	-1	79
21	1.328617e+19	1	14102100	1005	1	e8f79e60	c4342784	f028772b	ecad2386	7801e8d9	07d7df22	a99f214a	2fff7c47	90f22c00	1	0	20312	320	50	1780	0	1711	100077	48
22	1.726116e+19	1	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	c208bce7	0eb711ec	1	0	15704	320	50	1722	0	35	100084	79
23	3.878301e+17	1	14102100	1005	0	85f751fd	c4e18dd6	50e219e0	3d3198ca	2347f47a	0f2161f8	dc1c01c4	7e43a399	5096d134	1	2	18993	320	50	2161	0	35	-1	157
24	1.392136e+19	1	14102100	1005	0	1fbe01fe	f3845767	28905ebd	ecad2386	7801e8d9	07d7df22	a99f214a	a7e0eaf6	1f0bc64f	1	0	15701	320	50	1722	0	35	100084	79
25	1.768403e+19	1	14102100	1005	0	5b08c53b	7687a86e	3e814130	ecad2386	7801e8d9	07d7df22	a99f214a	b4983313	375c3d47	1	0	19015	300	250	2162	2	39	-1	33

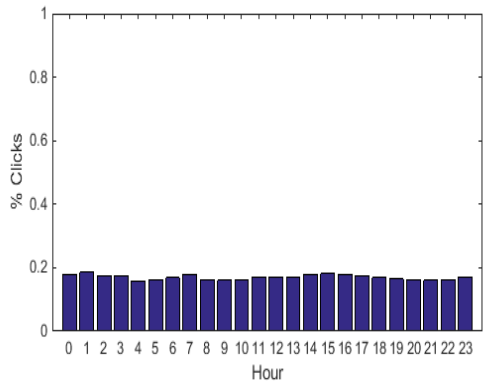
Displayed 1000 rows of 500,048 (499,048 omitted)

Big data or small computers?

- Training set = 10 days of data
- 5.87 GB comma separated values file (aka a text file)
- 40,428,967 observations with 24 columns!
- All variables are categorical
- Test set = 1 day of data
- 4,577,464 observations
- Personal computers faltered with the data
- Dual Xeon processors and 48 GB of RAM or Azure didn’t
- Down sampled for practicality

What the #@&%?

- Most variables were obfuscated
- Even missing values were obfuscated
- Sample values such as f3845767
- No units or descriptions on other values
- No information on bounds
- Are there proxies such as device ip and device id?
- Portion of data that clicked is constant for each hour


% Clicks
Hour
0
0.2
0.4
0.6
0.8
1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23

The final countdown

- Only 6,865,066 or 16.98% clicked the ad
- 240 unique hour long segments
- Many variables with large number of categories

Variable	Number of unique categories	Correlation Coefficient with click
'id'		
'click'	2	1
'hour'	240	-0.00778
'C1'	7	-0.03525
'banner_pos'	7	0.025535
'site_id'	4,737	-0.01604
'site_domain'	7,745	-0.02951
'site_category'	26	-0.01854
'app_id'	8,552	-0.04409
'app_domain'	559	-0.03002
'app_category'	36	-0.04555
'device_id'	2,686,408	
'device_ip'	6,729,486	
'device_model'	8,251	-0.00359
'device_type'	5	-0.03776
'device_conn_type'	4	-0.08636
'C14'	2,626	-0.06113
'C15'	8	-0.0946
'C16'	9	-0.13744
'C17'	435	-0.06197
'C18'	4	0.021634
'C19'	68	-0.03157
'C20'	172	-0.07076
'C21'	60	-0.07104

A chip off the old block

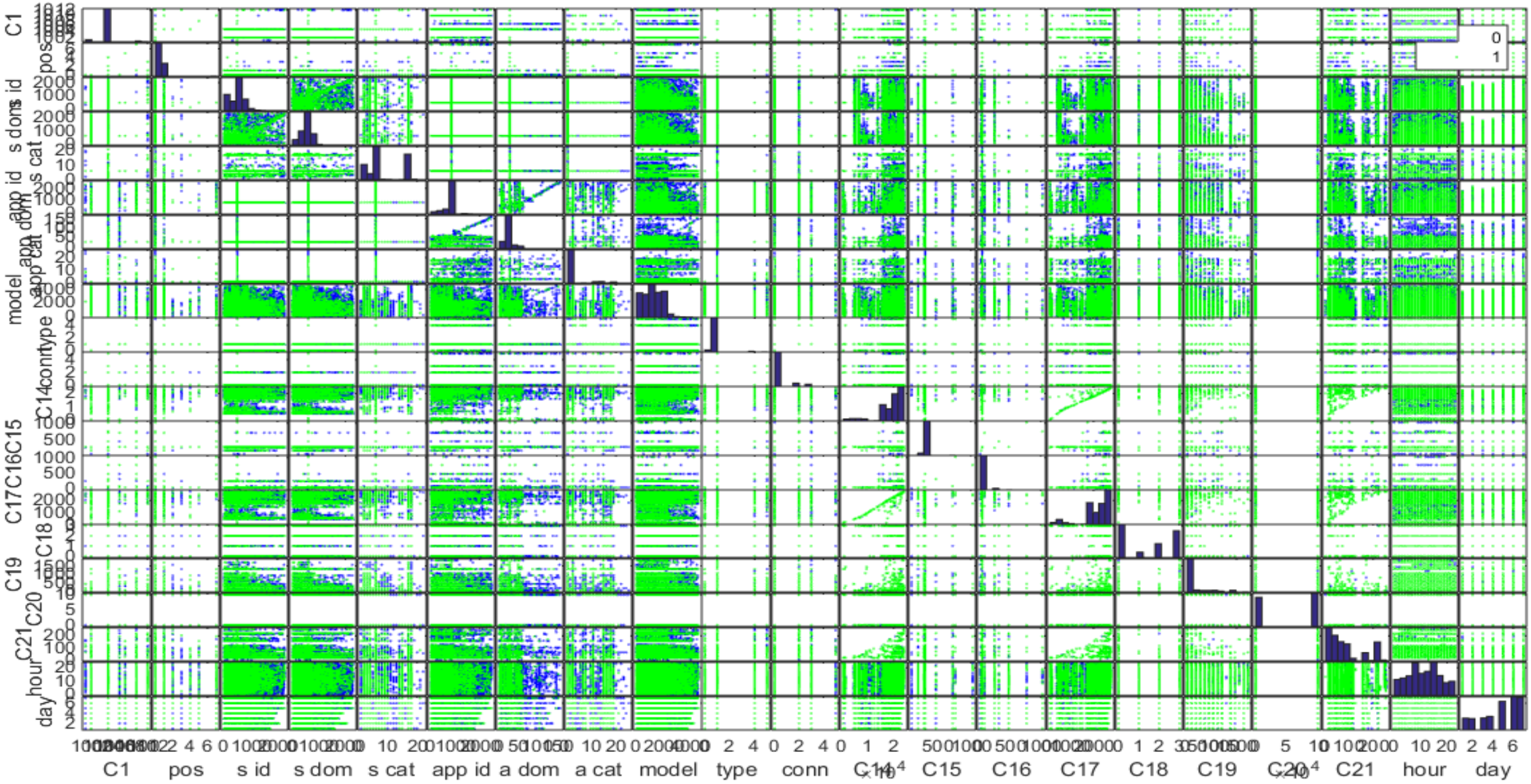
- Example sampling strategies
  - Stratified sampling to preserve the click ratio
  - Stratified to preserve proportion from each hour long segment
  - Multiple independent samples
- Code example-MATLAB
- Random observation
  - “Sampling 50/50 or 83/17? When using 50/50 sample with WEKA-classify-cross validation, I was not able to get ‘correctly classified instances rate’ above 65%. So decided to use the sample with 83/17 split which gets 80%+. I don’t really understanding why.“ Anita Miller

Feature transformations

- Eliminated suspected proxy variables
- Broke up the data/time variable to day of week and hour of day
- Binning or removal of some of the high categorical count variables
- Null values-to be or not to be?
  - Actually tells us something
  - Need to make predictions where there are missing values
- Attribute scaling?
  - Unsure, technically everything is categorical but…
- Can some variables be treated at numerical values instead of categorical?
  - C15 and C16 are believed to be the dimensions of the ad in pixels

Feature selection

- Feature selection using many methods
- Group plot matrix
- Correlation
- Gain ratio
- Information gain (via decision trees)
- Wrapper methods
  - ReliefF
  - SequentialFS (linear classifier + bagged decision trees)
  - CfsSubsetEval + BestFirst
  - CfsSubsetEval + Random Search
  - Infogain + Ranker
  - Chisquared +Ranker
  - Gainratio + Ranker
  - Classifier + Ranker
- Tallied votes and picked top performers


C1
pos
s id
s dom
s cat
app id
a dom
a cat
model
type
conn
C14
C15
C16
C17
C18
C19
C20
C21
hour
day
0
1

And the finalists are…

- Banner position
- Site domain
- Site category
- App domain
- App category
- Device conn type
- C14
- C16
- Hour of the day
- Day of the week

- Site id
- Site domain
- Site category
- App id
- App domain
- App category
- Device model
- Device conn type
- C14
- C16
- C18
- C21

- Site domain
- App domain
- C14
- C16
- C18
- C21

Models tried

- Logistic regression
- Decision trees
- Bagged decision trees/Random forests
- Boosted decision trees
- SVM
- Neural networks
- And more to come…

Name	Description	Feature Engineering	Score
Random guess benchmark	Random guessing with all predictions set to a probability of 0.5 of clicking the ad	None	0.6931472
Boosted Tree-10% data	Boosted decision tree from Azure ML with a maximum of 20 leaves per a tree, minimum of 10 samples per leaf, a learning rate of 0.2, the number of trees set to 100, and trained at the entire dataset.	None	0.3993284
SVM	SVM from Azure ML with only 10% of training data.	None	0.4352991
Boosted Tree-all data Model A	Boosted decision tree from Azure ML with a maximum of 20 leaves per a tree, minimum of 10 samples per leaf, a learning rate of 0.2, the number of trees set to 100, and trained on the entire dataset.	None	0.3990766

Name	Description	Feature Engineering	Score
Boosted Tree-all data Model b	Boosted decision tree from Azure ML with a maximum of 40 leaves per a tree, minimum of 100,000 samples per leaf, a learning rate of 0.2, the number of trees set to 100, and trained on the entire dataset.	None	0.4049064
Random 17% Benchmark	Matching population distribution in sample data, 17% of the test data was randomly selected with a uniform probability and given a 100% probably of clicking the ad, the remaining test data was given a 0% probability of clicking the ad.	None	9.5526002
Baseline Neural Network	Azure ML default parameters neural network with entire Kaggle dataset	None	0.4678139
Logistic Regression	Linear logistic regression from R, single predictor banner_pos, 500k sample set	Categorical	0.4411265

Name	Description	Feature Engineering	Score
Logistic Regression	Linear logistic regression from R, C14 and C16 with interactions only, 500k sample set	None	0.4361704
Logistic Regression	Linear logistic regression from R, C14 only, 500k sample set	None	0.4413473

Conclusions

- Garage in equals garage out
  - ‘Curated’ data files
  - Lacking variable descriptions
- What’s big?
  - There are a lot of numbers but it’s really not scary
  - Just need a little more memory than found in standard consumer computers
- Try, try, and try again
  - Our best score, so far, found out of naïveté
  - New submissions lead to new understanding

Tools of the trade

- MATLAB
- R
- Weka
- Python
- Alteryx
- Azure ML
- Excel