Todd Schultz

Team 3 Trees

Kaggle Project Homework – Due December 8, 2014

# Feature Selection Summary

Variables chosen to start with:

* banner position
* site domain
* site category
* app domain
* app category
* device conn type
* C14
* C16
* hour of the day
* day of the week

Site category and site domain are chosen to stand in for site id and app domain and app category is chosen to stand in for app id.

# Team Baseline

First, here is a public baseline set by Kaggle. Then a baseline of randomly choosing 17% to click.

### All 0.5 Benchmark, score = 0.6931472

Random guessing with all predictions set to a probability of 0.5 of clicking the ad

### Random 17% click, score = 9.5526002

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. So, unless you get really lucky and pick the exact right entries you’re going to fail.

### Team baseline, score = 0.3990766

Best score as of November 21, 2014 11 am.

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.

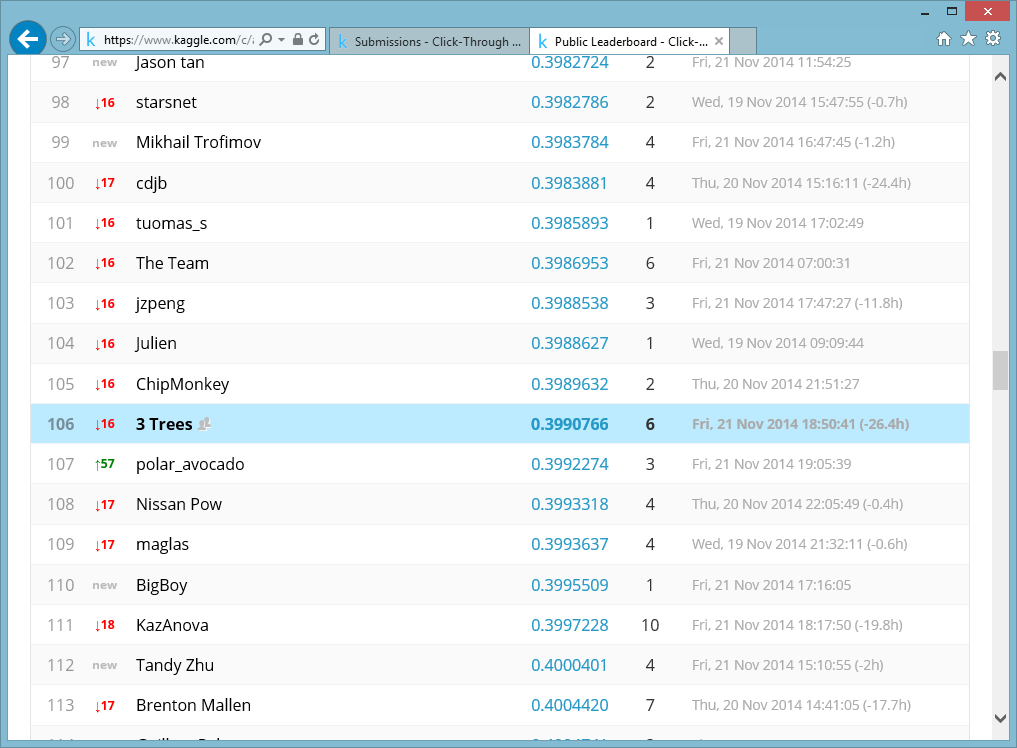


Figure 1: A sample of the Kaggle public leaderboard for Avazu Click Through Rate Challenge as of November 21, 2014 11:15 am.

## Model comparisons

|  |  |  |  |
| --- | --- | --- | --- |
| 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 |
| 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 |
| 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 |

### Comments

So far, our best model is performing almost twice as good as randomly guessing, which is encouraging. The boosted decision tree appears to perform better than the SVM but the SVM should be given a chance with optimized parameters. The smaller number of leaves per tree and the smaller number of minimum samples per leaf appear to perform better.

# Feature Engineering

## Original Dataset

The original training dataset provide by Kaggle is a 5.87 GB csv file with a total of 40,428,967 observations and 24 columns from 10/21/2014 12:00 am UTC to 10/30/2014 11:00 pm UTC. This file is deemed too large to work on most computers and thus will be resampled to reduce its size. The 24 columns are 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, and C21. All predictor variables are categorical. The hour variable is the time code segmented into hour long segments written as YYMMDDHH. The number of unique categories for each variable is given below along with the naïve estimate of the correlation coefficient with the click (response) variable. Nothing truly stands out as a ‘must have’ variable so more work is needed. Just on a hunch, the ‘hour’ variable was grouped into 24 unique levels, one for each hour of the day and the correlation recomputed. The correlation coefficient between clicks and this new hour variable was -0.0015.

Table 1: Correlation coefficients with respect to the variable 'click' for the entire dataset provide by Kaggle.

|  |  |  |
| --- | --- | --- |
| 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 |

To understand the population of the original data better before resampling, the proportions of the populations for two variables was investigated. Out of the entire 40,428,967 observations, only 6,865,066 observations clicked on the ad for a percentage of 16.98 %. Next, I determined that the number of observations for each hour long segment was not uniform. Both the percentage that clicked and the percentage in each hour long segment would have to be accounted for in the resampling strategy. The last thing considered before resampling was the percentage that clicked on the ad for each hour of the day. Again, here the hour variable was grouped into 24 segments, one for each hour of the day. The figure below shows the breakdown of the percentage that clicked for each hour of the day. The percentage of observations that clicked the ad for each hour of the day is uniform to within 3%, thus will be considered uniform. This leads me to believe that either the original dataset was sampled to enforce equal percentages or the hour variable will not be an important predictor.



Figure 2: Percentage of observations for each hour of the day that clicked on the ad.

## Resampled Dataset

The decision to resample the data and reduce its size was made for practical reasons. The original dataset provided by Kaggle is simply too large for most consumer grade (laptop) computers to handle in a time efficient manner. If I was working on this data as a business venture, I would look for computational resources to handle the data size such as a workstation class desktop computer or cloud resources. Here, resampling will increase the productivity of our team by allowing the use of their home, consumer grade computers. I generated two sample dataset with the same strategy, but with varied amounts of data. The smaller dataset was designed to have 100,000 samples and the second one was designed to have 500,000 samples. Both were also provided with a smaller, independent validation dataset so that the entire training set may be used for training the model. The sampling strategy was a two layered stratified sampling technique that preserves the percentages of those that clicked the ad and the percentages of data from each hour long segment, all 240 of them. The major steps are outlined below.

1. Determine overall percentage of clicks and data in each hour segment.
2. Initialize new training dataset with the first row of the original data and the validation dataset with the second row.
3. Pull in the next group of observations and the total number of observations are present and determine how many hour long segments are present.
4. Break the data sample into each hour long segment and then subdivide each hour long segment into those that clicked and those that didn’t click.
5. Pull a random sample without replacement proportional to the total number of samples desire to the number of observations in each subsample and save as the training set.
6. Remove all samples that have been saved into the training set from the subsample groups.
7. Pull another random sample without replacement proportional to the number of samples desires for the validation set and save.
8. Repeat for the clicked and no clicked data.
9. Repeat for each hour long segment.
10. Repeat until the end the original data file from Kaggle has been reached.

The new reduced size samples were then checked for consistency with the percentage of clicks and with the population distribution of data in each hour long segment and were found to be identical with regards to those two statistics. Now the reduced size datasets can be used for machine learning. The 500k sample contained 500,048 observations in the training set and 99.949 observations in the validation set. The minor difference in the final observation counts compared to the desired values were due to rounding to integer values for each data pull from the original file. The 100k sample contained 99,949 observations for the training set and 50,049 for the validation set.

## Feature selection and creation

From the work from the entire dataset that produced the correlations with the click variable and the investigation with the hour variable some conclusion can be drawn. The hour variable doesn’t look to promising in its current form or as a just the hour of the day. The variable C16 looks the most promising with the largest correlation coefficient. Before going any further the raw hour variable is transformed into two new variables, hourofday and dayofweek. My hope is that these variables together will be useful for the prediction. Also, the device\_id and device\_ip variables are strings with an enormous amount of unique categories and thus are difficult to deal with. I will ignore these variables in my preliminary feature analysis and reconsider them again later.

The remainder of this section will use the 500k sample exclusively and where all data has been converted to a numerical type for computations. My first step is to gain an overall understanding of the data by visualizing a scatter plot of each variable against each other and grouped by the state of the click variable, the predictor. The scatter plot shows some interesting features and can help identify variables that may be useful for model. In particular, the figure can also show where a combination of two variables might be suitable predictors. A visual inspections show potential for the following variables: C1, site id, site domain, app domain, device model, C14, C17, C19, C21, and hour of the day and day of the week when combined with other variables in the list. Next, the mean of each variable was computed for the two groups designated by the click variable. The variables with a percent difference greater than 10% are banner position, app category, device conn type, C16, C20, and C21. The correlation coefficients are computed next for the numerical data. The largest correlation coefficient with the response variable, click, was 0.13 with C16. Strong correlations greater than 0.9 were detected for two variable pairs. Variables C1 and device type had a correlation coefficient of 0.9 and variables C14 and C17 had a correlation coefficient of 0.98. This suggests redundant information I will plan to only keep one of variables from each pair for modeling later.



Figure 3: Scatter plot of 21 variables against each other grouped by the click variable.

Table 2: Mean values for each variable separated by the click variable state.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Response Variable | Click = 0 | Click = 1 |  | % difference |
| 'C1' | 1004.98 | 1004.88 |  | 0.01 |
| 'ban pos' | 0.28 | 0.32 |  | 11.36 |
| 's id' | 482.82 | 472.79 |  | 2.10 |
| 's dom' | 536.83 | 515.38 |  | 4.08 |
| 's cat' | 8.69 | 8.58 |  | 1.29 |
| 'app id' | 749.90 | 772.10 |  | 2.92 |
| 'app dom' | 26.87 | 26.59 |  | 1.05 |
| 'app cat' | 3.09 | 2.62 |  | 16.27 |
| 'model' | 1230.95 | 1229.36 |  | 0.13 |
| 'type' | 1.02 | 0.97 |  | 5.31 |
| conn type' | 0.36 | 0.18 |  | 69.07 |
| 'C14' | 18964.18 | 18236.04 |  | 3.91 |
| 'C15' | 319.22 | 317.46 |  | 0.55 |
| 'C16' | 57.30 | 73.16 |  | 24.31 |
| 'C17' | 2125.76 | 2044.57 |  | 3.89 |
| 'C18' | 1.42 | 1.49 |  | 4.69 |
| 'C19' | 228.35 | 225.00 |  | 1.48 |
| 'C20' | 54620.02 | 46881.36 |  | 15.25 |
| 'C21' | 85.02 | 73.69 |  | 14.28 |
| 'hour' | 11.28 | 11.28 |  | 0.00 |
| 'day' | 4.86 | 4.86 |  | 0.00 |

Next, two advanced methodologies were used to determine feature importance. The first methodology was the ReliefF algorithm and provides a rank assignment for each variable. The results for the 500k dataset are shown in the table below. The top five variables identified by the RefliefF algorithm in order are hour of day, C18, app id, app domain, and C15. The use of app id in the model will be difficult since that variable contains 8,552 different categorical values and isn’t guaranteed to contain all possible values. Next, a sequential feature selection algorithm is used both forwards and backwards for two different base models. The first base model was a simple linear classifier and the forward search selected only two variables, device conn type and C16. The backwards search selected the device model, C14, C16, and C17. The use of the device model variable is subjected to the same difficulties as the app id variable. The next model used for the sequential feature selection was a decision tree model. The forward search identified site id, app id, and C20, while the backwards search identified banner position, site category, app id, app category, C16, C17, C19, and day of the week. The last method used to identify features was based on the reduction of out-of-bag error when creating a bagged decision tree or random forest model. This was carried out first by constraining all the variables to be categorical variables and then by only allowing C1, C14, C15, C16, C17, C18, C19, and C21 to be numerical variables. The top five variables identified for the first run with all categorical variables are site id, app id, device model, hour of the day, and day of the week. The top five variables identified for the partial categorical list are device model, C14, C16, hour of the day, and day of the week.

Table 3: ReliefF feature selection results.

|  |  |  |
| --- | --- | --- |
| Response Variable | ReliefF Weights | ReliefF Rank |
| C1 | 0.0025 | 20 |
| banner position | 0.0124 | 9 |
| site id | 0.0589 | 12 |
| site domain | 0.0569 | 18 |
| site category | 0.0189 | 21 |
| app id | 0.0478 | 3 |
| app domain | 0.0215 | 4 |
| app category | 0.0189 | 6 |
| device model | 0.0984 | 15 |
| device type | 0.0017 | 17 |
| device conn type | 0.0211 | 7 |
| C14 | 0.0724 | 11 |
| C15 | 0.0039 | 5 |
| C16 | 0.0039 | 8 |
| C17 | 0.0315 | 19 |
| C18 | 0.0043 | 2 |
| C19 | 0.0217 | 16 |
| C20 | 0.0661 | 13 |
| C21 | 0.0127 | 14 |
| hour of day | 0.1059 | 1 |
| day of week | 0.0593 | 10 |

The results from all of the feature selection methods are assembled into a summary table presented below. A score of one was given to each variable selected by a certain method and then the scores were tallied across the methods to determine an overall score. The top scoring variable is C16, which is believed to one of the dimensions in pixels of the ad. Thus this variable could be represented numerically or as a categorical variable. The next highest scoring variables are app id, device model, hour of the day, and day of the week. The app id and device model variables are problematic for modeling as they have over 8,000 unique categorical values and are not guaranteed to contain all possible values. I believe that these variables could be engineered to a useful state if the true, unobfuscated values were available. I would say the same for the two variables that I have ignored until this point, device id and device ip, which I feel are pseudo-proxies for the identification number. But I could probably extract information from those variables with the real values such as the domain triplet. At this point, I’m going to move forward to modeling without using device id, device ip, app id, and device model. Instead, I’m going to use the following list of variables to start with and see from the modeling results which variables are needed. The variable I’m starting with are banner position, site domain, site category (site domain and site category will stand in for site id), app domain, app category (app domain and app category will stand in for app id), device conn type, C14, C16, hour of the day, and day of the week.

Table 4: Correlation coefficients for the 500k dataset.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | click | C1 | Ban pos | site id | site dom | site cat | app id | app dom | app cat | device model | device type | device conn type | C14 | C15 | C16 | C17 | C18 | C19 | C20 | C21 | hour of day | day of week |
| click | 1.00 | -0.04 | 0.03 | -0.01 | -0.04 | -0.01 | 0.03 | -0.01 | -0.04 | 0.00 | -0.04 | -0.08 | -0.06 | -0.03 | 0.13 | -0.05 | 0.02 | 0.00 | -0.06 | -0.06 | 0.00 | 0.00 |
| C1 | -0.04 | 1.00 | 0.29 | -0.05 | 0.02 | 0.03 | -0.09 | 0.00 | 0.09 | 0.06 | 0.90 | 0.19 | 0.06 | 0.12 | 0.06 | 0.07 | -0.04 | 0.00 | -0.04 | 0.04 | 0.01 | 0.02 |
| banner position | 0.03 | 0.29 | 1.00 | 0.28 | -0.36 | 0.53 | 0.14 | 0.03 | -0.22 | 0.05 | 0.32 | -0.08 | -0.01 | 0.06 | 0.03 | -0.03 | 0.10 | 0.13 | 0.05 | -0.10 | 0.00 | 0.00 |
| site id | -0.01 | -0.05 | 0.28 | 1.00 | -0.02 | 0.42 | -0.01 | 0.00 | 0.02 | 0.03 | -0.05 | 0.00 | -0.01 | 0.00 | -0.03 | -0.01 | 0.29 | 0.08 | 0.09 | -0.15 | 0.02 | 0.03 |
| site domain | -0.04 | 0.02 | -0.36 | -0.02 | 1.00 | -0.50 | -0.10 | -0.03 | 0.15 | -0.01 | 0.02 | 0.10 | 0.00 | 0.06 | -0.11 | 0.01 | -0.14 | -0.03 | -0.06 | 0.09 | 0.01 | -0.02 |
| site category | -0.01 | 0.03 | 0.53 | 0.42 | -0.50 | 1.00 | 0.14 | 0.05 | -0.21 | 0.02 | -0.01 | -0.14 | 0.00 | 0.00 | -0.11 | -0.01 | 0.15 | 0.11 | 0.11 | -0.12 | -0.01 | 0.03 |
| app id | 0.03 | -0.09 | 0.14 | -0.01 | -0.10 | 0.14 | 1.00 | 0.05 | -0.36 | -0.01 | -0.10 | -0.23 | -0.05 | 0.04 | 0.07 | -0.05 | -0.02 | 0.08 | -0.03 | 0.03 | -0.05 | -0.03 |
| app domain | -0.01 | 0.00 | 0.03 | 0.00 | -0.03 | 0.05 | 0.05 | 1.00 | -0.18 | 0.00 | 0.00 | -0.02 | 0.02 | 0.03 | 0.01 | 0.02 | -0.02 | 0.01 | 0.18 | -0.09 | 0.04 | -0.05 |
| app category | -0.04 | 0.09 | -0.22 | 0.02 | 0.15 | -0.21 | -0.36 | -0.18 | 1.00 | 0.00 | 0.08 | 0.23 | 0.00 | 0.04 | -0.06 | 0.02 | 0.08 | 0.02 | -0.06 | 0.06 | 0.03 | 0.06 |
| device model | 0.00 | 0.06 | 0.05 | 0.03 | -0.01 | 0.02 | -0.01 | 0.00 | 0.00 | 1.00 | 0.06 | 0.00 | -0.01 | -0.01 | 0.00 | -0.01 | 0.06 | 0.01 | 0.01 | -0.03 | 0.00 | 0.00 |
| device type | -0.04 | 0.90 | 0.32 | -0.05 | 0.02 | -0.01 | -0.10 | 0.00 | 0.08 | 0.06 | 1.00 | 0.21 | 0.05 | 0.18 | 0.07 | 0.05 | -0.05 | 0.00 | -0.05 | 0.04 | 0.01 | 0.01 |
| device conn type | -0.08 | 0.19 | -0.08 | 0.00 | 0.10 | -0.14 | -0.23 | -0.02 | 0.23 | 0.00 | 0.21 | 1.00 | 0.07 | 0.07 | -0.01 | 0.08 | -0.06 | -0.01 | 0.09 | 0.06 | 0.03 | -0.05 |
| C14 | -0.06 | 0.06 | -0.01 | -0.01 | 0.00 | 0.00 | -0.05 | 0.02 | 0.00 | -0.01 | 0.05 | 0.07 | 1.00 | 0.00 | 0.04 | 0.98 | -0.23 | -0.13 | 0.02 | 0.41 | -0.05 | 0.13 |
| C15 | -0.03 | 0.12 | 0.06 | 0.00 | 0.06 | 0.00 | 0.04 | 0.03 | 0.04 | -0.01 | 0.18 | 0.07 | 0.00 | 1.00 | -0.07 | 0.00 | 0.01 | 0.05 | 0.01 | 0.00 | 0.00 | -0.01 |
| C16 | 0.13 | 0.06 | 0.03 | -0.03 | -0.11 | -0.11 | 0.07 | 0.01 | -0.06 | 0.00 | 0.07 | -0.01 | 0.04 | -0.07 | 1.00 | 0.05 | 0.08 | -0.07 | -0.05 | -0.08 | 0.01 | -0.02 |
| C17 | -0.05 | 0.07 | -0.03 | -0.01 | 0.01 | -0.01 | -0.05 | 0.02 | 0.02 | -0.01 | 0.05 | 0.08 | 0.98 | 0.00 | 0.05 | 1.00 | -0.25 | -0.13 | 0.01 | 0.42 | -0.05 | 0.15 |
| C18 | 0.02 | -0.04 | 0.10 | 0.29 | -0.14 | 0.15 | -0.02 | -0.02 | 0.08 | 0.06 | -0.05 | -0.06 | -0.23 | 0.01 | 0.08 | -0.25 | 1.00 | 0.09 | 0.01 | -0.54 | 0.02 | -0.08 |
| C19 | 0.00 | 0.00 | 0.13 | 0.08 | -0.03 | 0.11 | 0.08 | 0.01 | 0.02 | 0.01 | 0.00 | -0.01 | -0.13 | 0.05 | -0.07 | -0.13 | 0.09 | 1.00 | 0.09 | -0.14 | 0.00 | -0.05 |
| C20 | -0.06 | -0.04 | 0.05 | 0.09 | -0.06 | 0.11 | -0.03 | 0.18 | -0.06 | 0.01 | -0.05 | 0.09 | 0.02 | 0.01 | -0.05 | 0.01 | 0.01 | 0.09 | 1.00 | -0.04 | 0.02 | -0.04 |
| C21 | -0.06 | 0.04 | -0.10 | -0.15 | 0.09 | -0.12 | 0.03 | -0.09 | 0.06 | -0.03 | 0.04 | 0.06 | 0.41 | 0.00 | -0.08 | 0.42 | -0.54 | -0.14 | -0.04 | 1.00 | -0.07 | 0.20 |
| hour of day | 0.00 | 0.01 | 0.00 | 0.02 | 0.01 | -0.01 | -0.05 | 0.04 | 0.03 | 0.00 | 0.01 | 0.03 | -0.05 | 0.00 | 0.01 | -0.05 | 0.02 | 0.00 | 0.02 | -0.07 | 1.00 | -0.02 |
| day of week | 0.00 | 0.02 | 0.00 | 0.03 | -0.02 | 0.03 | -0.03 | -0.05 | 0.06 | 0.00 | 0.01 | -0.05 | 0.13 | -0.01 | -0.02 | 0.15 | -0.08 | -0.05 | -0.04 | 0.20 | -0.02 | 1.00 |

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Response Variable | Plot Matrix | % Diff. of Mean | Correlation Coefficients | ReliefF | Sequential-linear-forward | Sequential-linear-backward | Sequential-tree-forward | Sequential-tree-backward | Random Forest-partial cat | Random Forest-all cat | Total Score |
| C1 | 1 |  | a |  |  |  |  |  |  |  | **1** |
| banner position |  | 1 |  |  |  |  |  | 1 |  |  | **2** |
| site id | 1 |  |  |  |  |  | 1 |  |  | 1 | **3** |
| site domain | 1 |  |  |  |  |  |  |  |  |  | **1** |
| site category |  |  |  |  |  |  |  | 1 |  |  | **1** |
| app id |  |  |  | 1 |  |  | 1 | 1 |  | 1 | **4** |
| app domain | 1 |  |  | 1 |  |  |  |  |  |  | **2** |
| app category |  | 1 |  |  |  |  |  | 1 |  |  | **2** |
| device model | 1 |  |  |  |  | 1 |  |  | 1 | 1 | **4** |
| device type |  |  | a |  |  |  |  |  |  |  | **0** |
| dev conn type |  | 1 |  |  | 1 |  |  |  |  |  | **2** |
| C14 | 1 |  | b |  |  | 1 |  |  | 1 |  | **3** |
| C15 |  |  |  | 1 |  |  |  |  |  |  | **1** |
| C16 |  | 1 | 1 |  | 1 | 1 |  | 1 | 1 |  | **6** |
| C17 | 1 |  | b |  |  | 1 |  | 1 |  |  | **3** |
| C18 |  |  |  | 1 |  |  |  |  |  |  | **1** |
| C19 | 1 |  |  |  |  |  |  | 1 |  |  | **2** |
| C20 |  | 1 |  |  |  |  | 1 |  |  |  | **2** |
| C21 | 1 | 1 |  |  |  |  |  |  |  |  | **2** |
| hour of day | 1 |  |  | 1 |  |  |  |  | 1 | 1 | **4** |
| day of week | 1 |  |  |  |  |  |  | 1 | 1 | 1 | **4** |