When an ad is displayed to a user online, it is referred to as an impression (see http://en.wikipedia.org/wiki/Impression\_(online\_media)). This dataset contains a portion of about a month’s worth of impressions for one campaign, focusing on user interactions with the ads. An interaction is counted if the user moves her mouse cursor over the ad. Each impression has an image associated with it, called a "creative". Creatives can be of different size (for example, a popular banner ad size is 728x90 pixels).

We define the interaction rate (IR) for a set of impressions as the number of impressions with at least one interaction divided by the total number of impressions.

cm\_impression.csv - impression-level data

-----------------

impression\_id - unique impression id

any\_interaction - whether or not the user interacted with the ad (1 or 0)

user\_id - integer representing a unique user

date\_offset - data was collected on consecutive days, starting on "day 0"

day\_of\_week - a string representing the day of the week

hour - the hour of the day (0 to 23)

site\_id - an integer id of a particular site where the ad was displayed

creative\_id - an integer representing a unique creative for the campaign

ad\_size - size of the creative (image) in pixels

cm\_user.csv - user-level data

-----------------

user\_id - unique integer representing a user

zip\_id - an integer representing the user's zip code

browser\_id - an integer representing the user's web browser

connection\_type\_id - an integer representing the user's internet connection type

cm\_geo.csv - geographical data

-----------------

zip\_id - an integer representing a zip code (note: this is not the zipcode)

state - 2-letter abbreviation of the state for this zip code

region - a string representing the geographical region for this zip code

Questions

1. Which ten states have the most unique users who have impression data?

NY, TX, FL, NJ, PA, CA, GA, NC, IL, MD

-----------------------------------------------------------------------------------------------------

2. How many unique users who were served impressions don’t have geographic information? 130728

What is their interaction rate? 0.0404657

-----------------------------------------------------------------------------------------------------

3. Which five sites have the highest interaction rate? Site\_id= 65, 114, 52, 225, 116

Give a reasonable confidence estimate for the interaction rate for these sites. How would you pick five sites with the highest interaction rate to put in a research report for the advertiser, and why?

I used bootstrap method over “any.interaction” to define the confidence intervals except, when the number of observation is equal to 1 or all the observations are the same (which happens usually when the number of observations is very low).

Based on this criterion the confidence interval for the first two site ids (i.e., 65 and 114) is not defined. Then I use 95% confidence interval to pick the site ids with higher ineraction rates, and here is the result:

####

site.id int\_rate bs\_lower\_CI bs\_higher\_CI

52 0.4465409 0.4353703 0.4860466

225 0.4339731 0.4302337 0.4376384

116 0.3806147 0.3490156 0.3908589

215 0.3500690 0.3467903 0.3515207

111 0.3213115 0.2972621 0.3599004

#####

-----------------------------------------------------------------------------------------------------

4. What is the distribution of the number of impressions and the interaction rate by:

- the hour of the day (consider weekends and weekdays separately)

- the day of the week

- browser

- connection type

- geographical region

- the size of the creative

I plotted all distributions …

-----------------------------------------------------------------------------------------------------

5. Are there any significant differences between geographical regions by the hour? How would you visualize this?

The frequencies of hours in “South” and “Northeast” region are higher compare to “Midwest” and “West” region.

I depicted this difference by plotting region vs. hours where the size of each point in each (hour,region) is proportional to the frequency of (hour,region). I used qplot {ggplot2} for plotting.

-----------------------------------------------------------------------------------------------------

6. Tell us something about the data that you find interesting and that wasn't covered in a previous question. Show us why this is interesting with a visualization and explanation.

First of all there is pattern which is common in all regions: the frequency start to increase for all regions from hour 7-8 to 18-19 and then it decrease from 19 to 23 and it increase again at 0 (or 24) and it finally decreases from 0 to 7-8.

I also normalized frequency within each region and it seems that: although the overall frequencies of hours in “South” and “Northeast” region are higher compare to “Midwest” and “West” region but the rate of increase or decrease in the frequencies are other way around.

I showed the plots of Q5 and Q6 in one file for comparison purposes.

-----------------------------------------------------------------------------------------------------

7. Consider a model to predict interaction rate. Specifically, the model should be able to find impressions that are most likely to result in an interaction. Provide a summary of the steps in building the model. Explain your choice of the modeling method. Implementation of the model is optional, but you should at least process the data to prepare the input to the model. Describe how you will train your model and evaluate its performance.

Since the interaction rate is a number between 0 and 1 and we are looking for the probability of interaction, I choose a logistic regression model. Also I choose any.interaction as criterion variabale(i.e., y) and the rest as predictors.

I start with glm and lasso method to find the best set of predictor in the model. I order to tune lambda in Lasso I calculate the –fold cross valudation for a grid of lambda values. Then I chose the lambda with lowest cross vaidation error. Finally I refit the model with the selected lambda on the whole data.

-----------------------------------------------------------------------------------------------------