**Vericast Marketing Data Scientist – Analytical Exercise**

This assessment is designed to allow you to showcase your analytical abilities on a simplified version of a problem we usually work on.  You will have an opportunity to walk a couple of the Marketing and Data Scientists through your solution over a video call. This session will be a structured, but informal, discussion regarding the distributions found in the data and conclusions you draw from them.

We don't expect this to be a time-intensive exercise for you. At a minimum, we expect you to answer the two questions below fully and to the best of your ability, with all code and output shown, well formatted, and well documented. You can choose any widely adopted language (Python, R, Julia, ...) and tool (Jupyter notebooks, Rmd, ...) to best showcase your work. However, we do have a preference for Python in a Jupyter notebook because it's what we use internally.  Please have any code used to generate solutions or output you think useful in describing your approach ready to present during your presentation.

**Problem Description**

Vericast has run a digital advertising campaign for one of our clients, targeting mobile users in the Southeast. The goal of the campaign was to drive conversions -- users clicking through the ad and accepting our client's offer.

Our campaigns are executed on real-time bidding (RTB) exchanges. Web pages or mobile apps submit bid requests to potential advertisers on these exchanges. If Vericast sees a bid request for a user we want to target on a web page or mobile app we want to serve on, we will submit a bid for that advertising placement. If we submit the highest (or only) bid, we win the placement and serve an ad. Given that we have won and served the ad, a user then has the opportunity to convert.

We ran two creatives (distinct ad designs) during the campaign: a baseline creative in line with our client's previous campaigns, and a new creative designed to increase conversions by more prominently displaying the offer to the user.  To do this, we assigned some users to a "test" group that received the new ad during the campaign.

**Data**

There are two files [available at this link](https://www.dropbox.com/sh/jzbvwceo2i7sy52/AAB8Ekwf6DtOMdaQSuc8lLS4a?dl=0) for download.  These files are as follows:

* *user\_attributes.csv.gz* is a compressed CSV file of target users' attributes
  + *user\_id* is a unique identifier for each user
  + *attributes* is a text blob of the user's attributes
    - *age* is the user's age in years
    - *gender* is the user's inferred gender
    - location is the user's state of residence
    - *test* is a binary variable indicating that the user belongs to the test group
* *bid\_requests.csv.gz* is a compressed CSV file of bid requests that we have seen for the targeted users. Each row represents one (bid) request.
  + *timestamp* is the time the bid request was received. This is unique for each request and can be treated as a unique key.
  + *user*\_id is the unique identifier for user the request is for
  + *bid* is a binary variable indicating whether we made a bid for this request
  + *win* is a binary variable indicating whether we submitted the winning bid and served an ad
  + *conversion* is a binary variable indicating whether the user converted after seeing the ad

**Exercises / Questions**

1. What do the different distributions of requests by user tell us?

Please describe the distributions related to requests and the users we targeted. We are interested in both visual and numerical analysis/explorations of the data distributions.

1. Based on insights from #1, was the test effective? Were test users more likely to convert than control users? *Did the new creative increase conversions?*

When you complete this exercise please contact whoever sent it to you so you can present your work, insights learned from your analysis, and conclusions pertaining to the questions asked.