Data Challenge

Nielsen Marketing Cloud - Data Science Team

Introduction

Thank you for your interest in joining the NMC Data Science Team! If you are reading this it means that we liked your CV and you have probably even made it through the first couple of interviews. We would now like to see how you conduct data science in practice. This is where the data challenge comes in.

We have made accessible to you a sample of actual production data. Find the instructions under the “Preparation” section of this document to help you get set up, then use the “Tasks” section as a loose guideline on exploring the dataset. Ideally this should take about 5 to 6 hours but the task list is non-exhaustive and you are encouraged to do extra if you find it interesting. Share the results with us in the way that you think is most appropriate, but we would like to see a prediction on the submission file (see below) and the final version of your code, be it in a notebook, a markdown document, or a github repository. We also ask you to treat the data and this task as confidential.

Thank you, good luck, and have fun!

Preparation

We provided you with four files:

- `data.csv` contains raw data with `user_uids` and associated “segments” (variables that represent what we know about these users). Each segment is identified by an ID and you will find a corresponding label in the `taxonomy.csv` file. The segments are also marked as P (when a user had the opportunity to click on an option in the quiz) and A (when the user actually clicked on an option in the quiz). Some segments are represented by binary variables, others by continuous variables.

- `taxonomy.csv` gives category labels for the variables in `data.csv`. Labels have an intrinsic hierarchy delimited by § characters.

- `conversion.csv` represents conversion data we may receive from a client. These are customers that we know have purchased a product of interest in the past. This csv contains the information on the users who converted. The rest of the users in the data.csv file (except the users in the submission file for which we want to predict) can be considered as nonconverters.

- `submission.csv` lists the `user_ids` that we would like predictions for.

Okay up to here? Great, let's go!

Tasks

1. Take a quiz: http://you.visualdna.com/quiz/whoamidc#/quiz. The data you are going to analyse was produced by a quiz very similar to this one. Print the results as a PDF (make sure to check the option for saving the background images so that the actual scores get stored in the resulting PDF). Keep it for later, we would like a copy to get to know you better!

2. Get the data we provided into R or Python. The choice is yours! Have a look at the data. The score column contains a `JSON` like structure with keys and values. Transform the table into a format that makes these values accessible for further analysis. Give us the breakdown of the prevalence (non-zero occurrence) of each variable (key), e.g.:

Label	Count
10153_A	4264
10153_P	7493
10157_A	682
10157_P	7494

3. Conduct an exploratory data analysis on the dataset and share your insights with us. Use them to clean the dataset as you see fit before proceeding to the next step.

4. Let's say our client is a carribean holiday operator. They would like to advertise their trips to potential customers online. They've provided a sample of users that we know have bought tours from them in the past, which are saved as `user_uids` in `conversion.csv`. We would like to know which of the other users in our dataset are likely to be interested in their offers as well.

   Create a model using an algorithm of your choice to predict Caribbean holiday interest and populate the field `p_conv` in `submission.csv` with predictions for the respective `user_uids` given in the file.

5. Send your quiz results (task 1), the `submission.csv`, your analysis code, and any supporting information such as plots, reports, presentations etc back to us.

We expect you to ask some questions during this task. Please email them to muktamala.chakrabarti@nielsen.com. Good luck with the task!