PayPal Sales Analytics Data Science Take-Home Case Study

Instructions

Directions (time window, submission type, submission method)

- We are providing 3 days for you to work on the case study
- Our intention is that you would spend 3 hours or less actively working on it
- Use Python or R in a Jupyter or Rstudio notebook for your analysis
- The notebook should contain an easy-to-find answer to the two questions in the problem statement
- You will submit the Notebook **and** a PDF or html version through an email to your recruiting contact
- The team will grade your submission using a standard rubric and will discuss the case study with you during one of the interview sessions

Rules

- You may use whichever packages you want.
- You may consult general online or text resources (e.g., Google "how to do X in Python", coding manuals, etc.) without attribution.
- You may re-use any self-created or open source notebook templates.
- You may re-use any self-created code snippets, classes, functions (e.g., a scikitlearn pipeline setup).
- You may adapt any found code snippets (e.g., from a Medium article or a StackOverflow question) *provided that you cite the source*. If you go this route we *will* evaluate the trustworthiness of the source when grading your submission.
- You **may not** re-use, adapt, or share this case study with friends, colleagues, online, etc.

Advice

There are many acceptable ways to solve this problem (i.e., you can use some type of machine learning if you'd like but a heuristic solution can be rated just as highly). Many valid solutions do not use every single attribute provided–we want you to be judicious in deciding what is worth investigating as your time = money. Our intention is to let you use your preferred methods to arrive to a "good enough" conclusion, not to have you

comprehensively characterize the data or create executive-ready charts. We are definitely not trying to trip you up with obscure character sets, erroneous data, etc.

Case study

Motivation and context

A vendor representing large US media publishers contacts you offering a list of subscribers to their flagship landscape architecture monthly magazine[1]. The vendor has provided sample data which they state has been randomly selected from an active subscriber base of >30,000 accounts.

One of your job duties is to evaluate new data sources for generating sales leads, so it falls on you to do a quick analysis to decide if you should bring the vendor data to your boss's attention. Your boss has high trust in you and you are encouraged to filter out any "hard no" vendor offerings yourself. The price quoted for the vendor data is in line with norms and is within budget so the benefit:cost ratio is not a concern; your boss is more interested in the potential to generate new revenue from the leads.

In performing your assessment, you should keep in mind that the Home & Garden vertical has been growing quickly and is a current sales focus area for PayPal. The available data indicates that PayPal penetration into this space is relatively low, so leaders are very interested in enlarging the top of the sales funnel. An addition of 1000 high-quality leads would count as a meaningful increase.

In your experience, a critical determinant in the value of such data sources is their alignment with one of PayPal's target merchant audiences. The expected annual revenue generated by a qualified lead from those target merchant audiences are[2]:

- large independent online sellers who do not offer PayPal: $70k-$200k
- medium businesses with online or invoice sales: $10k-$40k
- small businesses on channel partners (in particular, Shopify, Magento, and BigCommerce): $1k-$5k

Leads outside of those groups are not valuable to the sales team. For example, casual sellers or customers who use PayPal to shop online or transfer money to friends and family generate $0-$200 of revenue annually.

[1] A subscription offers both print and online access

[2] These are not real numbers!

Problem statement

Your objective is to perform a first-pass exploratory data analysis (3 hours or less) and evaluate (1) whether the vendor offering is worth discussing with your boss and (2) why or why not.

Included files

- A sample CSV/TSV (`subscriber_data_sample`) from the vendor listing email addresses, industry, length of relationship with magazine (years), and total number of site visits to the magazine's website
- A CSV/TSV (`pp_cust_data`) with email addresses of PayPal users and account send / receive active status (active = has transaction within last year)

Givens

- The anonymization converts any name or other personally identifying information into an equal-number of upper case letters.
- H&G or business-related words are not anonymized; numbers and punctuation have also been left as-is.
- The anonymization is consistent across both data sets (e.g., if `maflint@paypal.com` becomes `AGSAMLQ@paypal.com` in the PayPal data it would also be translated to `AGSAMLQ@paypal.com` in the vendor data).
- The colleague who pulled the internal data did a good job and the set provided would capture all accounts likely to also exist in the vendor data.

Evaluation

We care about:

- Your rationale for the approach taken. We will discuss your reasoning during the interview but it will be helpful if you provide comments in the Notebook explaining your choices.
- Whether the approach taken is reasonable given the motivation and context.
- The extent to which your decision on whether and why to (not) discuss the vendor data with your boss is backed up by your analysis.
- The extent to which you accomplished (or added to!) a list of "good choices" compiled by the authors of this assignment. The "good choices" are compiled across a variety of reasonable problem solving approaches so no submission is expected to cover everything).

- Notebook structure (use of cells to organize your analysis in a logical fashion).
- Code hygiene and interpretability.

We do not care about:

- Which programming language and packages (if any) you choose.
- Resource-intensiveness or optimality of any packaged or user-defined algorithms.
- The level of "polish" of text and graphs as long as someone with an analytics or data science background can interpret them.
- Specific coding style (indentations, spaces, variable naming conventions, etc.).
- The number of insights delivered past the minimum needed to come to a conclusion.