Background:

At Scowtt, our core objective is turning messy client data into actionable scores. We do this by building robust transformation pipelines that clean and featurize disparate datasets, allowing our machine learning models to accurately predict conversion probability and rank order target audiences.

Take Home Assessment:

In a real world scenario, you will rarely start with a perfect dataset. You will have to aggregate data from different sources to create a clean dataset to train our models. We will provide .csv files describing the user's journey through an online sales platform. The main ones are:

- customers_dataset.csv (Lead): All of the potential customers
- order_items_dataset.csv (Browsing): The items they added to their cart
- orders_dataset.csv (Purchase): When an order is placed

Typically the sales cycle is Lead -> Browsing -> Purchase. **When data aggregating, make sure to aggregate the data to capture the user's journey in this fashion, merging event by event.** Ensure you thoroughly analyze the tables + merges, so that you can explain your methodology of data exploration/aggregation

Your Instructions

The files attached contain mock examples of raw customer data. Your task is as follows:

- **Data Exploration:** Dive into the provided files to understand the schema, identify key identifiers, and surface any data quality issues.

- **Data Aggregation:** Merge and transform these sources into a single, clean master dataset. (Please refer to the schema, if needed). Again, our goal is to predict on a *per user basis*, so aggregate it in a way that the model can learn/predict per user

- **Model Training:** Use your aggregated dataset to train a predictive model of your choice to identify high value advertisement targets.
  - Output should be a predictive ***'Score'*** (0-1 scale) for the propensity of a user to place an order in a specific

timeframe, let’s say 30 days, as well as that user’s predicted ***conversion value (Order value)***

  - You can create one model can produce the two outputs above, or two separate models for each of the required outputs

- **Evaluation:** Evaluate how well your model can score each user for their propensity to place an order. Provide metrics on model performance, score distributions, etc.

What to have in time for the interview: Complete this task in a python .ipynb notebook with all the code cells executed with final results.

What to send before the interview: Your completed .ipynb notebook with the output results so we can examine it. Be ready to explain your code + results. You can also send a github link if that works better.

Notes:

- **Process Documentation:** Briefly explain your rationale for your data cleaning and feature engineering choices.
- You should not have to spend more than 2-3 hours on this

- For the model architecture you are free to make your own design decisions (number of layers, activation function, etc.). Optimize for accurate and efficient results.

- You may not be familiar with some of these technologies, so it is alright to use search, follow tutorials or use ChatGPT.

- Please do not use ChatGPT/or other AI tools to generate the whole code.

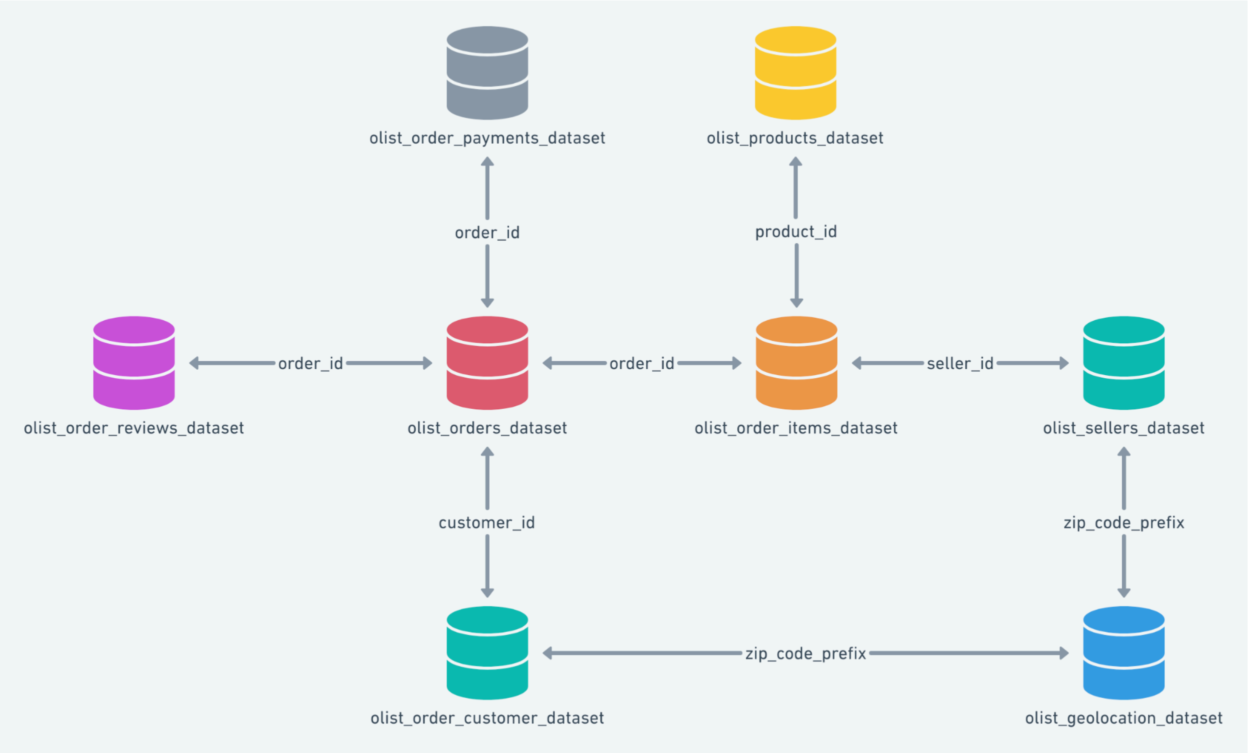
- If you are using ChatGPT/other AI tools, please share your session and prompts that you used.

- Please make sure to test your code properly.

  During a phone interview.

  - Our expectation is that you should be able to explain all the code and concepts used in the code in detail.

Schema:



Link to dataset: https://github.com/Scowtt-Inc/scowtt-shared/blob/bea550f17c084ad462f6d9c711960c687074750b/data/datasets.zip