



[EXTERNAL] Take Home Assignment - ML (Enrichment)

✓ 3 more properties

Background

Our goal with this homework assignment is to understand how you approach data enrichment problems from a practical and analytical standpoint. We're interested in seeing how you tackle complex, multi-faceted problems and synthesize information to propose effective solutions. Keep in mind that we are a remote, async culture that prefers written communication over presentations. You will be assessed both on your proposed solution and how well you communicate it.

Problem statement

At Monarch, providing our users with precise and actionable financial insights is a core objective. One significant challenge we face is transforming raw financial transactions into meaningful and accurate data. When users connect their financial accounts, we leverage secure data aggregation services like Plaid, Finicity, and MX to fetch transaction details. These services provide raw transaction data, which contains essential fields such as descriptions (like `#000000130058 DILLONS FUEL #7707 E. CENTRAL`), amounts (like `$5.99`), and dates (like `06/07/2024`). However, they often fall short in identifying accurate or complete meta-data regarding merchants (like `Dillon's Fuel`) and categories (like `Gas & Transportation`).

This gap in data quality presents a problem: without correct merchant names and categories, users cannot fully understand or manage their finances effectively. Misidentified or uncategorized transactions can lead to confusion, misinformed budgeting, and inefficient financial planning. For instance, a vague transaction entry might fail to inform the user whether a charge was for "Netflix" or another service, and an inaccurate category might misrepresent an "Entertainment" expense as a "Gift", potentially impacting that user's budgeting for the month.

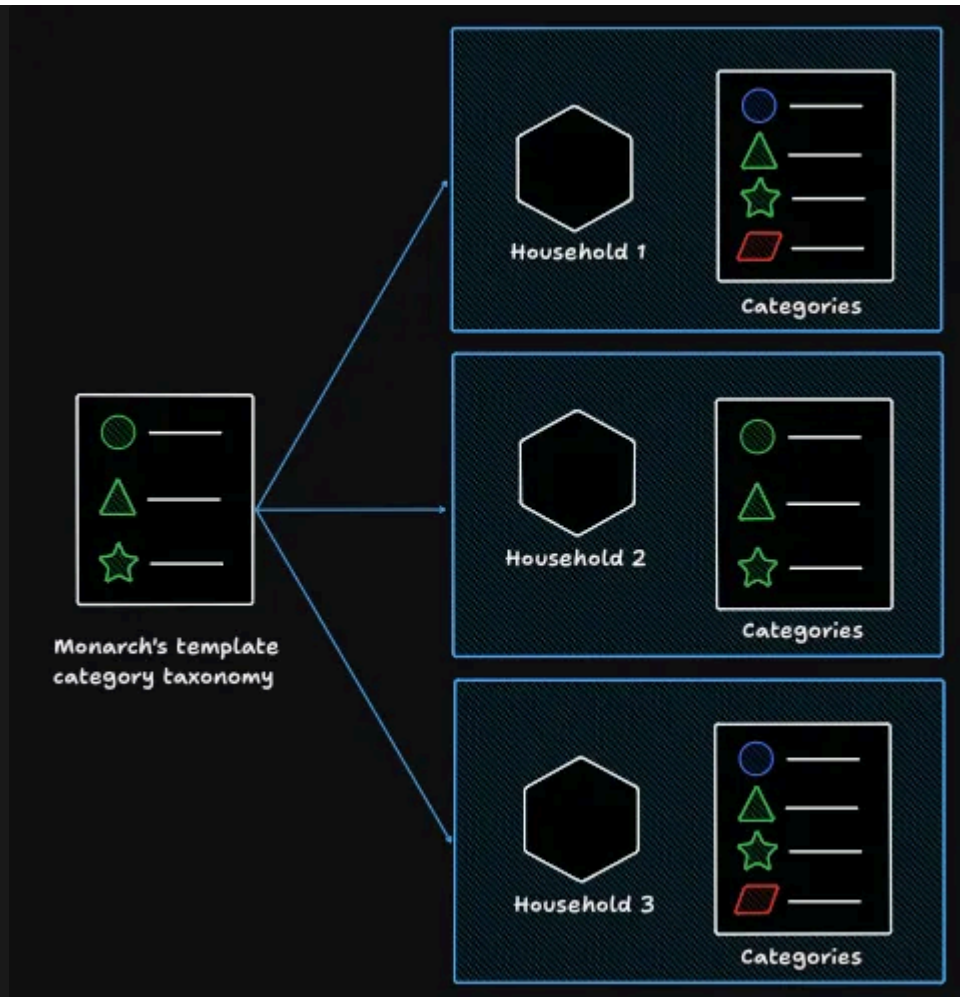
To address this, Monarch must implement robust transaction enrichment processes. This involves accurately determining the merchant and category for each transaction using the available data. The goal is to convert basic transaction details into rich, precise information that aligns with our taxonomy, enabling users to have a clear, organized, and comprehensive view of their financial activities.

Transaction enrichment is therefore vital for ensuring our users have reliable data to make informed financial decisions. By tackling this challenge, Monarch can enhance user experience, facilitate better financial management, and ultimately empower users to achieve their financial goals with confidence.

Category enrichment

Let's assume that **category enrichment** in Monarch, i.e. determining the category associated with a given transaction, is currently achieved by a simple look up of static mappings between the aggregator's category name, present in the raw transaction data, and a category in a user-specific category taxonomy, in a process that takes place when ingesting/syncing transactions for a maximum of ~20 transactions at a time.

The user-specific taxonomy utilized in this lookup process is initialized upon household creation and copied from a template taxonomy defined by Monarch. Any categories derived from this initial template are referred to as "system categories". System categories can be renamed and extended by users, but they can't be deleted. Users can add custom categories at any time.



As an example of a category enrichment, assume we get a transaction from, say, Finicity, containing:

```
{ "description": "THE MUSEUM OF MODERN ART (MOMA) TICKET #1234",
  "data_aggregator_category": "Arts" }
```

Where the field `description` contains a raw string returned by the financial institution from which our data aggregator is pulling the transaction data, and `data_aggregator_category` describes the transaction's category according to the aggregator's taxonomy.

Assume also that our Finicity to system category mappings contain:

```
FINICITY_CATEGORY_MAPPINGS = { "Advertising":  
    SystemCategory.ADVERTISING_PROMOTION, "Air Travel":  
    SystemCategory.TRAVEL_VACATION, "Alcohol & Bars":  
    SystemCategory.RESTAURANTS_BARS, "Allowance":  
    SystemCategory.FUN_MONEY, "Amusement":  
    SystemCategory.ENTERTAINMENT_RECREATION, "Arts":  
    SystemCategory.ENTERTAINMENT_RECREATION, # <----- ... }
```

we assign the category `SystemCategory.ENTERTAINMENT_RECREATION` to the incoming transaction.

In order to collect data that could help us improve category enrichment, we also implemented a mechanism to record every manual category assignment hoping that we can use this data to improve the current solution. Basically, whenever a user assigns one of their transactions to a different category than the one Monarch assigned using the technique above, we save both the old and the new data, e.g.:

```
{ "transaction_description": "THE MUSEUM OF MODERN ART (MOMA)  
TICKET #1234", "old_category": "Entertainment & Recreation",  
  "new_category": "Hobbies" }
```

Based on the context in this document, we would like you to propose an alternative solution that leverages the available data to improve **category enrichment**. Please make sure you account for the following aspects:


- You can assume that each raw transaction contains the `description` and `data_aggregator_category` fields;
- Each of the data aggregators we integrate with has their own category taxonomy;
- Each Monarch household has their own set of categories to which you have access. When we create a new household, it is instantiated as a copy of our system categories (i.e. every household is created with the same set of categories). However, users can add new categories and rename existing ones. System categories cannot be deleted, only renamed;
- Given that category enrichment takes place as part of the transaction ingestion background task, the less latency it adds to the process the better;


- We usually ingest transactions in small batches. You can assume these batches do not exceed 20 transactions and belong to the same household (to which you have access);

Task

Inputs

- Transaction dataset

 enrichment_evidence_2024-07-08T12_24_19.655026+00_00.csv 1070.5KB

 categories_2024-07-08T12_24_19.655026+00_00.csv 83.7KB

- Default categories

- See [this](#)

► Default category mappings for Plaid

Task I: Analysis + Design

In this task you will write a document that

1. Formally defines the problem we want to solve;
2. Performs a preliminary data exploration to characterize the available dataset and understand its nature as well as that of the underlying process;
3. Proposes a model or a combination of models that can be used to solve the problem given our requirements and the characteristics of the data available, assessing the pros and cons of this solution;
4. Proposes how the model can be trained (what are the tools you would use, what kind of infrastructure would be required, how often will we train the model, what are the associated costs, etc) and evaluate it.
5. Propose how the model can be deployed and maintained (required infrastructure, observability, costs, etc)

You are not expected to actually train any models or implement anything. Instead, imagine that this document is an initial design document or Request For Comment meant to outline possible approaches, gather feedback from your teammates, and guide any future implementations.

Feel free to use whatever structure/format you deem most effective. If there is anything else you think could be important given the context in which this document will be used, make sure you include it. Likewise, if you identify any gaps in the requirements/specifications, please do point them out in the document, possibly also proposing what could be done to close them.

Deliverable: Data exploration + Design document