Mercari BI analyst/Data Scientist Exercise

Thank you for your interest in joining Mercari and taking the time to be part of our process. The insights and impact you'll have will help us grow our marketplace. Being able to identify patterns from data, provide recommendations and help business partners make informed decisions would be the key to success in this role.

Question 1 | Business Knowledge

In the first week of March 2018, Mercari marketplace experienced a spike in GMV (Gross Merchandise Value) and the business partners asked you to help analyze the potential drivers of this event. Provided you have all the necessary data, please list the data points you would check to complete your analysis.

For spike in GMV:

From a business perspective, good practice to already have in place tracked notable business events.

And from a technical perspective, it is good to consider system glitches, data integrity issues.

Given there is a specific date listed, good to reference Mercari US' history and see if there was a notable business event in the first week of March 2018. Already checked worldwide events during that time frame and nothing big happened in e-commerce.

Usually, when there are spikes in GMV they would come from notable business events (internal and external) such as the following:

1. Product Launches
2. Feature Releases
3. Marketing Promotions
4. Press
5. Competitor's Behavior (Amazon Prime Day)
6. Social media influencer caused viral purchase of particular collectible/beauty product
7. Fraud (surge in bot traffic)
8. Onsite Bugs
9. Analytics Bugs

Though after doing some research, in March 2018 Mercari US rebranded and changed the look of their mobile app, logo, icon, app UI and website. This helps highlight as well as reduce likelihood of notable business events that could have caused the spike in GMV.

With that new insight, in order to analyze potential drivers, data points I would check at business and tech levels would be:

Business-

1. Marketing team - confirm if rebranding happened during first week of March 2018
2. Marketing team - confirm if a spike in sales due to a product that is not one of the highest-selling Mercari items (tech, collectibles and home) due to a social media influencer raving about beauty product
3. Marketing team - confirm if Amazon Prime Day in 2018 was in first week of March 2018
4. CTO - ask if there was a new business initiative to use ML to leverage all data collected from sellers and buyers to create product recommendations for customer retention/more purchases, and if ML was used to provide price recommendations to sellers for them to sell their products faster
5. Analytics team - determine if there was a suspicious bot account with nonsensical location coordinates such as in the middle of the ocean
6. Engineering team - determine if there was a glitch in tracking website data that could have caused unusual spike in GMV
7. Engineering team - ask if something is wrong with data ingestion
8. Analytics team - determine if something went wrong during ETL process, reporting; check the whole data pipeline

Checking the data pipeline for root cause would involve:

Asking Data Engineers and Analytics Engineers-

1. During the first week of March 2018, was there some scheduled data audit that went wrong?
2. Is data in the middle of migrating warehouses?
3. Was there a non-updating table?
4. Did some tables never make it through data ingestion?
5. Was there a bad merge?
6. When looking at loads and null destinations and validations, were there any alarming findings?
7. When looking specifically at: Orders, Product and Users tables in DAG (Directed Acyclic Graph), were tables that these specific tables were dependent upon compromised in integrity?

In order to analyze and determine the root cause of a data anomaly, it is good to keep in mind:

1. It is possible that some of these events happened at the same time (rebranding and influencer's statement on a beauty product).
2. It is good to investigate what happened at the end of February 2018 to see if any notable events occurring during that time frame would lead to spike in GMV (chain of events).

And with that, going through the following process is beneficial:

A. Identify possible contributing factors (internal and external) - already done

B. Sort Factors (from highest likelihood causes of change all the way down to lowest possible cause of change)

- Brand redesign
- Fraud

When looking at GMV at multiple dimensions, see which segments contributed to spike. (Total GMV compared to GMV in beauty product/total GMV in LA

a. Which segments changed as the overall GMV changed? And of these which have the largest contribution to GMV? (e.g. GMV in LA trend as overall GMV and GMV in LA spike as overall GMV)

C. Classify Factors

Classify factor into one of four groups:

- Correlated Result
- Unrelated Factor
- Contributing Factor
- Root Cause- This is the factor that initiated the chain of events that resulted in the change. There may be more than one!

First step is to arrange everything in a timeline. Sometimes need to rely on knowledge of business and own organization's internal processes

Root Causes Analysis Best Practices:

1. Record Your Actions (keep track of significant business decisions and actions on shared calendar/spreadsheet)
2. Track External Forces (Monitor all external forces that might affect business

(competition, economics, governmental policy))

3. Segment Data (segment metrics to effectively evaluate likelihood of any given segment contributing to overall change)
4. Map Your Processes (business processes should be written down, so that we can map out difference between contributing and root cause event)

Question 2 | SQL Knowledge

Using the 3 tables below, please answer the following questions:

1. Top 3 product categories in GMV (sum of price) from last month
2. Categories that have GMV (sum of price) > $1,000,000 from last month
3. Weekly percentage of new registered customers who purchased within 7 days of registration
4. Daily average order size (GMV / total orders) by new customers vs. returning customers? [New customer is defined as customer purchased on the same day as registration]

Table Name: *orders*

Column Name	Data Type	Description
buyer_id	integer	ID of the buyer
seller_id	integer	ID of the seller
product_id	string	ID of the product
timestamp	datetime	Timestamp of the order
price	float	Price of the product

Table Name: *product*

Column Name	Data Type	Description
product_id	integer	ID of the product
category_id	integer	ID of the product category
brand_id	integer	ID of the product brand

category_name	string	name of the product category

Table Name: Users

Column Name	Data Type	Description
user_id	integer	ID of the user
created	timestamp	Timestamp of user registered
device	String	Device that user registered on (ios, android, web)

Answers to questions are at this link. In order to run the queries in the right panel, ‘Query SQL’, make sure your cursor is located in that panel. Then press ‘Run’ located to the right of the database version. SQL query results will display at the bottom.