Name: Isaac Ng Sheng

Disclaimer:
Uncertainty of which Dataset/Topic I should focus on. I have a general interest in Financial/Fintech Product Management and am considering a Credit Card Fraud Detection Dataset, however, the Amazon Sales Dataset has a more manageable number of attributes/features at 16 attributes (as compared to the Credit Card Fraud Detection Dataset having 122 attributes along with a second, previous_application csv).
I would love to hear your opinions on either Datasets (and which you would recommend for this project and for potential use in Financial/Fintech Product Management)

Topic(s)

Credit Card Fraud Detection Dataset:
https://www.kaggle.com/datasets/mishra5001/credit-card/data
Amazon Sales Dataset:
https://www.kaggle.com/datasets/karkavelrajaj/amazon-sales-dataset

Introduction(s)

Credit Card Fraud Detection Dataset:
Goal is to identify the causal link between the various attributes provided, against the probability of the individual defaulting on the loan. Possibly identifying strong indicators of default to build the company's portfolio and risk assessment.

Dependent Variable:
Target: Binary Variable with 1 being Default

Important Independent Variables (Non-Exhaustive, totalling to 121 Variables):

- Name_Contract Type: Identification if loan is cash or revolving
- Gender: Gender of the client
- Flag_Own_Car: Flag if the client owns a car
- Flag_Own_Realty: Flag if client owns a house or flat
- CNT_CHILDREN: Number of children the client has
- AMT_INCOME_TOTAL: Income of the client
- AMT_CREDIT: Credit amount of the loan
- AMT_ANNUITY: Loan annuity
- NAME_EDUCATION_TYPE: Level of highest education the client achieved
- FLAG_MOBIL: Did client provide mobile phone (1=YES, 0=NO)
- FLAG_EMAIL: Did client provide email (1=YES, 0=NO)
- OCCUPATION_TYPE: What kind of occupation does the client have
- OBS_30_CNT_SOCIAL_CIRCLE: How many observation of client's social surroundings with observable 30 DPD (days past due) default

Amazon Sales Dataset:
Goal is to identify the causal link between the various attributes provided, against the rating of the product. Possibly identifying strong indicators of what would lead to the greatest impact on the rating of the product.

Dependent Variable:
rating - Rating of the Product

Independent Variables:

- product_id - Product ID
- product_name - Name of the Product
- category - Category of the Product
- discounted_price - Discounted Price of the Product
- actual_price - Actual Price of the Product
- discount_percentage - Percentage of Discount for the Product
- rating_count - Number of people who voted for the Amazon rating
- about_product - Description about the Product
- user_id - ID of the user who wrote review for the Product
- user_name - Name of the user who wrote review for the Product
- review_id - ID of the user review
- review_title - Short review
- review_content - Long review
- img_link - Image Link of the Product
- product_link - Official Website Link of the Product

Methodology

The general process would be to perform Regression Analysis, possibly with Stepwise Model selection to maximize explainability (or R^2) with the least amount of variables to reduce dimensionality before deciding to add certain variables as controls for possible omitted variable bias.
Other means of ensuring internal validity would be to conduct IV Regression and Regression Discontinuity.
I would envision the IV for the Credit Card Fraud Detection Dataset to involve Cnt_Children to be used as the IV for Amt_Credit (which could be endogenous) as they tend to be relatively randomized thus, improving instrument exogeneity towards other attributes, while having instrument relevance to the Amt_Credit (as higher number of children tend to lead to higher fees and thus possibility for higher loan).
For the Amazon Sales Dataset, I would envision the discount_percentage could be a good IV for discount_price as it has instrument relevance to discount_price, being a causal link while being relatively exogenous to other variables.
To verify the effectiveness of the IV, I would conduct:

- Weak Instrument Test: Check the F-statistic from the first stage regression. (Ideally, F-statistic should be greater than 10)
- Overidentification Test: Attempt to attain more than one IV before conducting such test.

In the area of Regression Discontinuity, I would hold a threshold value, for the Credit Card Fraud Detection Dataset, it would be the Amt_Credit to set such threshold. Whereas for the Amazon Sales Dataset, the threshold could be the discount percentage or price.
If the situation allows, I would consider Difference-in-differences as a Causal Inference technique as well.

Finally, with the Regression Analysis (to be conducted after each specific attribute is chosen, arguing for the internal validity of the econometrics model), I would verify if each independent variable is statistically significant and then compare their coefficient to the dependent variable to determine the possible impact it could have.

Expected Contribution

Credit Card Fraud Detection Dataset:

1. Risk Assessment: Upon establishing a causal link between defaulters and the various attributes, the development of a model to predict loan defaults would be more attainable, enhancing the risk management strategies of financial institutions.
2. Policy Recommendations: Provide insights into which client attributes are strong indicators of default, aiding in the formulation of better credit policies.
3. Portfolio Management: Assist companies in building a more resilient loan portfolio by identifying high-risk clients early, and thus making more informed decisions.

Amazon Sales Dataset:

1. Marketing Strategies: Provide insights into which product attributes (e.g., discount percentage, category) most significantly impact ratings, aiding in the development of targeted marketing strategies for each specific product attribute.
2. Product Development: Offer recommendations for product improvements based on attributes that impact with higher ratings (assuming a causal link is established).
3. Sales Optimization: Help optimize pricing and discount strategies to maximize product ratings and sales to gain greater profit margins for the company..