DATA SCIENCE FRAMEWORK REPORT

DongMei Li, Ph.D.

May 2021

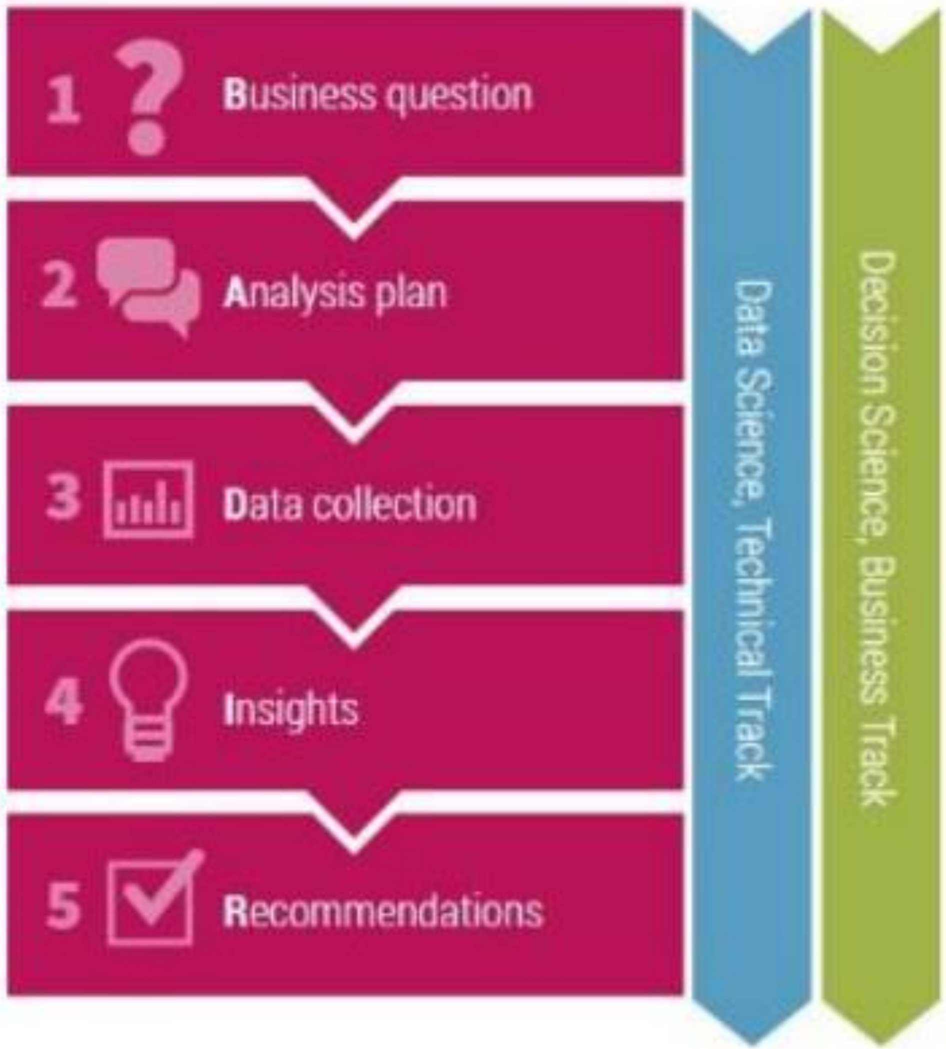
1. GOAL STATEMENT

- Enlisted by Credit One, our Data Science team will help design and implement a creative, empirically sound solution to the business-critical problem, loan default.

- The goal is to identify a much better way to understand how much credit to allow customers to use or, at the very least, if a customer should be approved or not.

2. DATA SCIENCE PROCESS FRAMEWORK

- The BADIR Framework is our data science process framework.
  - It makes more sense and it is important to craft the "analysis plan" (step 2) prior to the data collection.
    - In so doing, relevant data will be collected and prepared for analysis.
    - Coupled with step 1, successful returns on the analysis will be achieved.



THE BADIR FRAMEWORK

- Step 1: Business question. Credit One has seen an increase in the number of customers who have defaulted on loans they have secured from various partners, and Credit One, as their credit scoring service, could risk losing business if the problem is not solved right away.
  - We have been given full authority to solve this problem with whatever tools and methods we need.
- Step 2: Analysis plan
  1. Analysis goals: obtain the CreditOne data, check the data, do cleaning, EDA, model building and evaluation
  2. Hypothesis: test the relationship of demographic and payment related variables with default
  3. Methodology: how much credit to allow customers to use determines that this is a regression problem.
  4. Data specification: CreditOne data stored in MySQL database.
  5. Project plan: prepare and explore the data (2 weeks); build and evaluate models (2 weeks); present findings and recommendations (2 weeks)

THE BADIR FRAMEWORK, CONTINUED

- Step 3: Data Collection
  - Since CreditOne data stored in MySQL database, SQL is used to retrieve the data into a Pandas dataframe.
  - Data will be checked to see if there are duplicates and missing and all is numerical. Only relevant data are saved as .csv file for analysis.
- Step 4: Derive Insights
  - Review patterns in data
  - Prove or disprove hypotheses
  - Present findings in terms of quantified impact to guide prioritization of the hypotheses for analysis.
- Step 5: Recommendations
  - Engage the audience by presenting a short, concise, insightful set of recommendations without getting bogged down in detail.
  - Be perceived as an effective business partner by presenting credible recommendations.
  - Drive the audience towards actions that create impact by solving the business problem

3. DESCRIPTIONS AND LOCATION OF RELATED DATA SOURCES

- CreditOne data stored in MySQL database.
- SQL is used to retrieve the data into a Pandas dataframe.
- They are written as a .csv file in the task folder and can be imported (or to excel) for analysis.

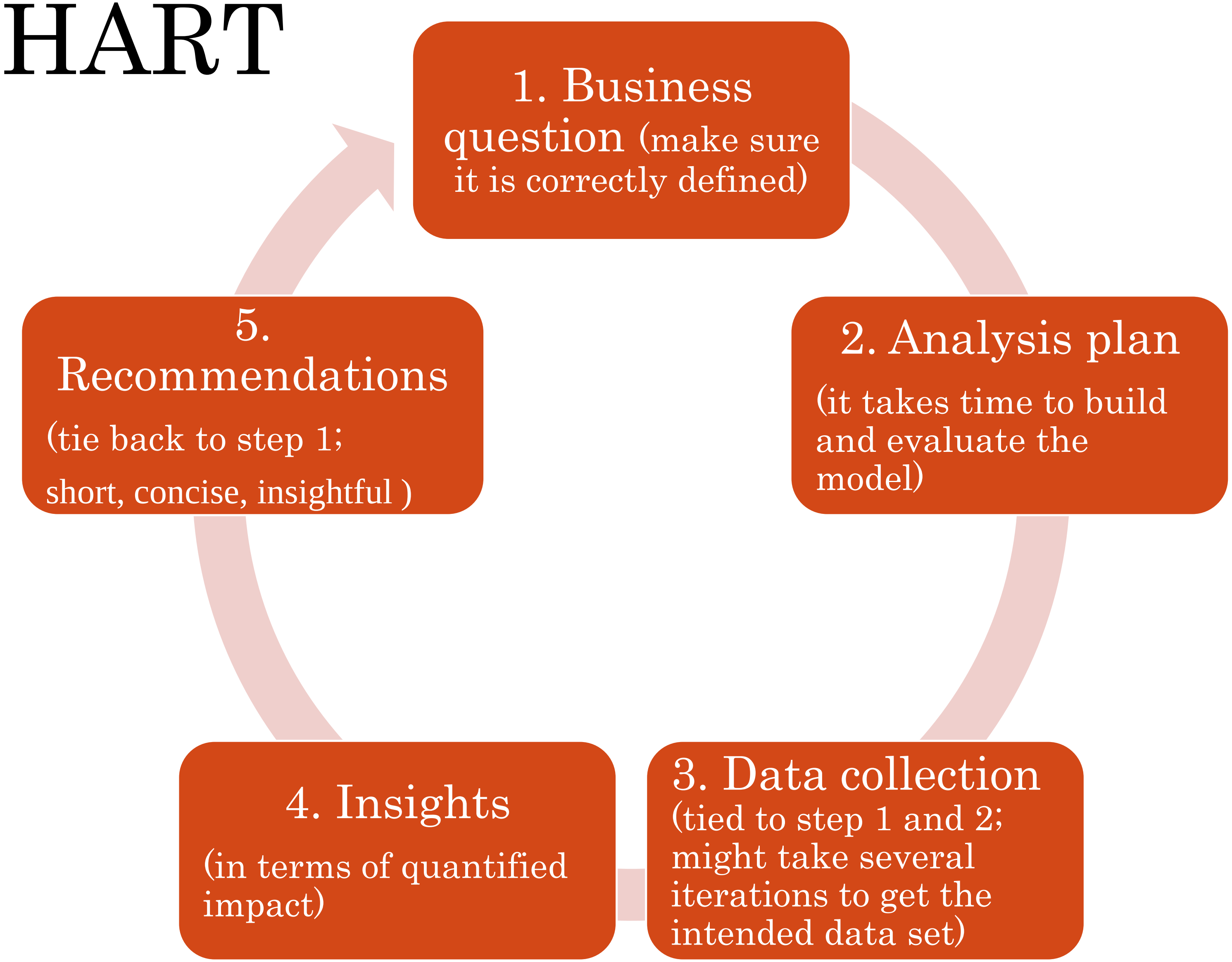
- According to the data source pdf document
  - Data source: Taiwan
  - There are 24 variables
    - 4 demographics (gender, education, marital status, and age)
    - amount of the given credit
    - 6-month history of past payment
    - 6-month amount of bill statement
    - amount of previous payment (for 6 months)
    - Binary default payment (Yes = 1, No = 0)

DATA MANAGEMENT & ISSUES

- 4. Data management:
    - SQL data that are retrieved and processed are written as .csv and excel and stored in the task folder for the team to access.
    - The .csv or excel data will directly be imported for analysis.

- 5. Any known issues with the data:
    - There are data duplicate and data type issues.
        - Weird values are displayed when plotting variables. Exploring csv data shows there are duplicates and unnecessary rows, which will be identified and removed in Jupyter Notebook.
        - Non-numerical data will be taken care of to prepare data for regression analysis.

6. FLOWCHART



1. Business question (make sure it is correctly defined)

2. Analysis plan

(it takes time to build and evaluate the model)

3. Data collection (tied to step 1 and 2; might take several iterations to get the intended data set)

4. Insights

(in terms of quantified impact)

5. Recommendations

(tie back to step 1;

short, concise, insightful)

7. INITIAL INSIGHTS

- There are some duplicates in the SQL data.
- More than 20 features in the dataset. Some are likely to have strong correlations and others have weaker relationships with the target/Y. Some weaker variables might have to be excluded.

- Data collection quality in the future can be improved.