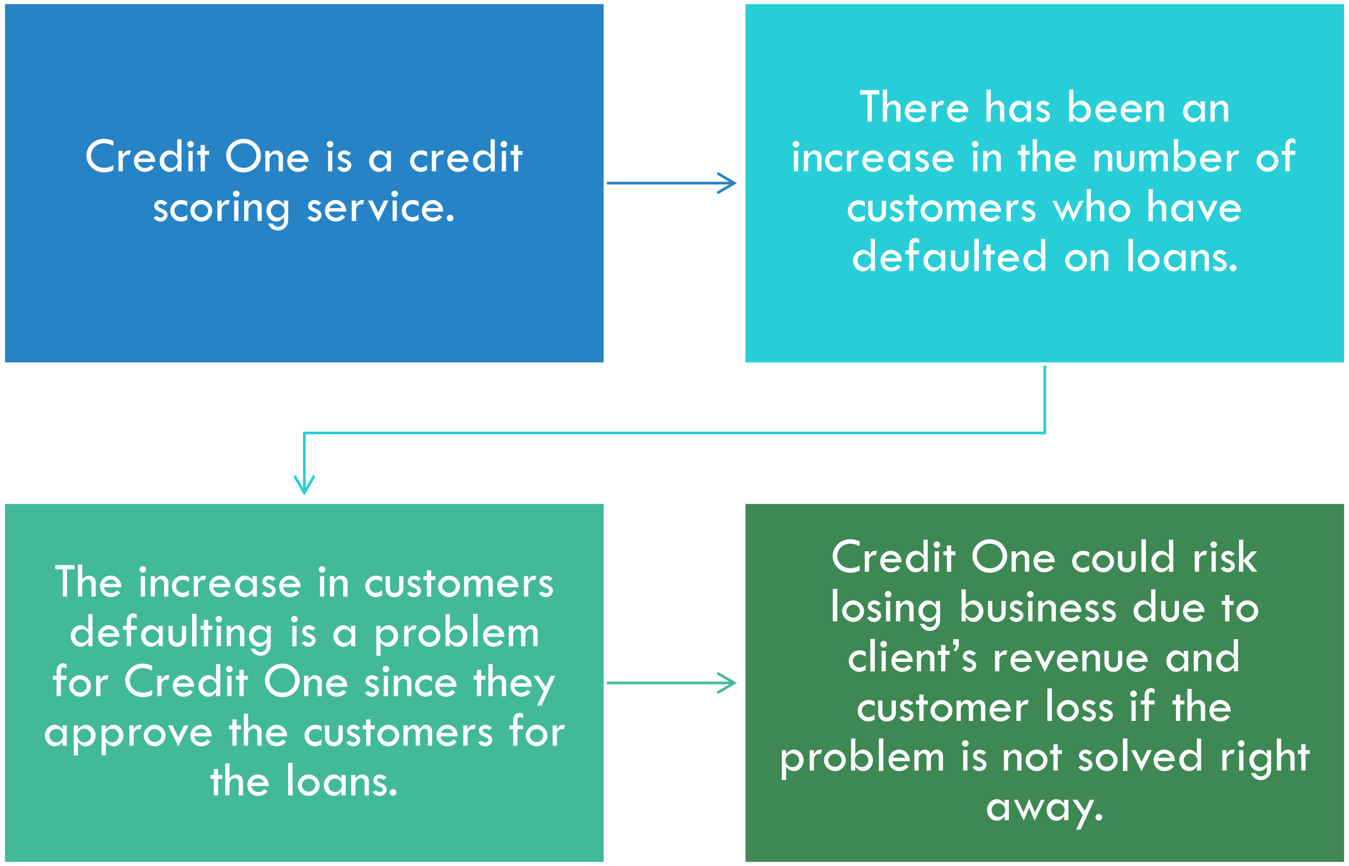
DATA SCIENCE FRAMEWORK FOR CREDIT ONE

Susana Reyes

BACKGROUND



Credit One is a credit scoring service.

There has been an increase in the number of customers who have defaulted on loans.

The increase in customers defaulting is a problem for Credit One since they approve the customers for the loans.

Credit One could risk losing business due to client's revenue and customer loss if the problem is not solved right away.

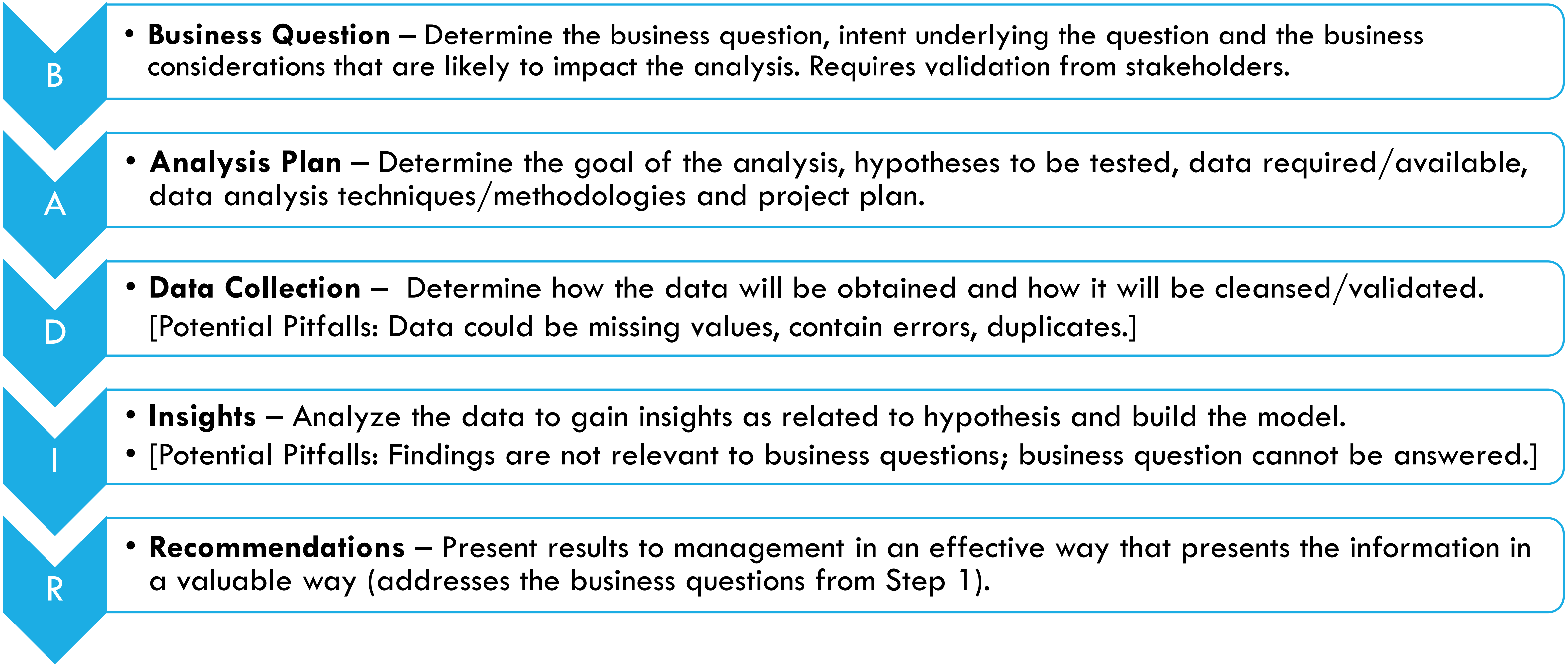
GOALS AND BUSINESS QUESTIONS

- **Goals:** Credit One needs a better way to understand how much credit to allow someone to use or, at the very least, if someone should be approved or not. The goal is to reduce the number of customers that default on their loans by creating a model that can indicate of a customer is likely to default or not on their loans.
- **Preliminary Business Questions:**
  - Should there be a limit to how much credit certain customers can borrow? What should that limit be?
  - Are certain customers more likely to default on their loans?
  - Are there certain customers we want to keep attracting and marketing to?
  - How do you ensure that customers can or will pay their loans? Is it possible to do this?

DATA SCIENCE FRAMEWORK

- A data science framework is essential to successful data analytics; every is step important and should be followed carefully.
- B.A.D.I.R is the acronym for the data science framework developed by the authors of *Behind Every Good Decision*, Priyanka Jain and Puneet Sharma.
- It stands for **B**usiness question, **A**nalysis plan, **D**ata collection, **I**nsights, and **R**ecommendations.
- The 5-step approach is easy to understand and follow.
- For models to be effective and relevant to the business, analysts need to revisit the framework to help them make the revisions and update the model.

FRAMEWORK PROCESS



B

- **Business Question –** Determine the business question, intent underlying the question and the business considerations that are likely to impact the analysis. Requires validation from stakeholders.

A

- **Analysis Plan –** Determine the goal of the analysis, hypotheses to be tested, data required/available, data analysis techniques/methodologies and project plan.

D

- **Data Collection –** Determine how the data will be obtained and how it will be cleansed/validated. [Potential Pitfalls: Data could be missing values, contain errors, duplicates.]

I

- **Insights –** Analyze the data to gain insights as related to hypothesis and build the model.
- [Potential Pitfalls: Findings are not relevant to business questions; business question cannot be answered.]

R

- **Recommendations –** Present results to management in an effective way that presents the information in a valuable way (addresses the business questions from Step 1).

DATA AND DATA SOURCES

Feature	Description
LIMIT_BAL	Amount of the given credit (NT dollar). It includes both the individual consumer credit and his/her family (supplementary) credit.
SEX	Gender (1 = male; 2 = female)
EDUCATION	Education Level (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others)
MARRIAGE	Marital Status (1 = married; 2 = single; 3 = divorce; 0=others)
AGE	Age
PAY_0 to PAY_6	History of Past Payment. Tracked the past monthly payment records / repayment status from April to September 2005 (-2 = No consumption, -1 = Paid in full, 0 = The use of revolving credit, 1 = payment delay for one month, 2 = payment delay for two months, 8 = payment delay for eight months, 9 = payment delay for nine months and above)
BILL_AMT1 to BILL_AMT6	Amount of bill statement (NT dollar). Tracked amount of bill statement from April to September 2005.
PAY_AMT1 to PAY_AMT6	Amount of previous payment (NT dollar). Tracked amount paid in April to September 2005.
Default Payment Next Month	Client behavior (Y=0 not default; Y=1 default)

Source: Credit One MySQL database containing 30,204 observations

DATA MANAGEMENT

- Data will only be used for stated purpose
- Data will only be stored for as long as required
- Data will be held in compliance with data protection laws and regulations
- Data will be secured against unauthorized or unlawful access and processing
- Data will be protected against accidental loss, destruction or damage
- Data will only be transmitted in encrypted form

DATA ISSUES

- Data contained an unnecessary header row
  - Dropped header row when importing CSV file
- Data contained duplicate rows
  - Dropped 204 duplicate rows
  - Remaining rows/observations: 30,000
- Data types were in the incorrect data type
  - Converted data types to be able to fully utilize them in model
- Data contained non-numeric features
  - Converted non-numeric features to numeric to be able to fully utilize them in model

ANALYSIS PLAN: EDA

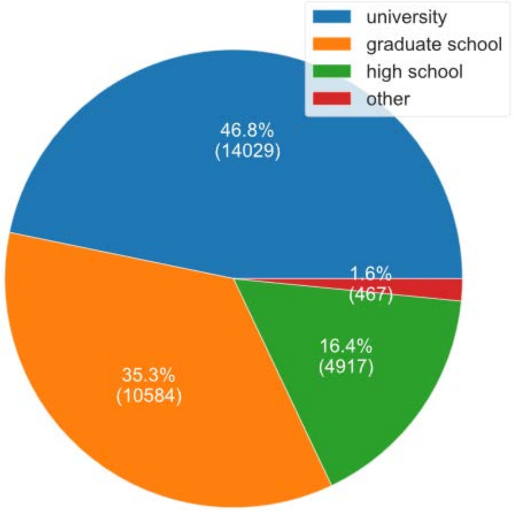
- Perform an extensive Exploratory Data Analysis (EDA) to answer the following questions:
  - What attributes in the data can we deem to be statistically significant to the problem at hand? Create histograms and scatterplots to better visualize the data, patterns, correlation, etc.
  - What concrete information can we derive from the data we have?
  - What proven methods can we use to uncover more information and why?
- Depending on EDA and data mining results, it could be necessary to revisit the business questions and analysis plan.

ANALYSIS PLAN: BUILD THE MODEL

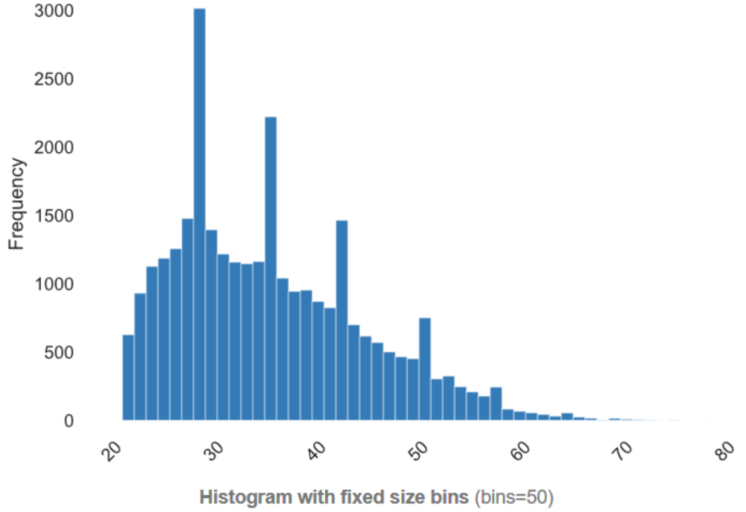
- We will be using machine learning predictive models, to start we need to:
  - Discretize certain features to be able to utilize them when modeling the data (for example, we can create bins for [Limit_Bal] and [Age]) .
  - Select the features to use as the dependent variable in our models.
    - Use [Default] as the dependent variable to investigate if we can predict what demographic of customers are likely to default.
    - Use [Limit_Bal] after discretization to investigate if we should limit the credit balance of certain customers.
  - Identify features that we should exclude from our models.
  - Run several different models, switching up the dependent variable to see if we can make any predictions.

INITIAL INSIGHTS

- Larger percentage of female customers (60.4%) v. male customers (39.6%)
- Larger percentage of highly educated customers (University: 46.8%, Graduate School: 35.3%, High School:16.4%, Other: 1.6%)
- Larger percentage of single customers (53.2%) v married customers (45.5%)
- Average age: 35. Median age: 34
- No missing values identified



Education Level



Age Histogram

NEXT STEPS

- Consult with stakeholders to confirm business questions.
- Collect data and ensure initial issues are resolved.
- Carry out analysis plan; revise as needed.
- Based on model, determine findings and insights as they relate the business questions.
- Present findings and recommendations to management.