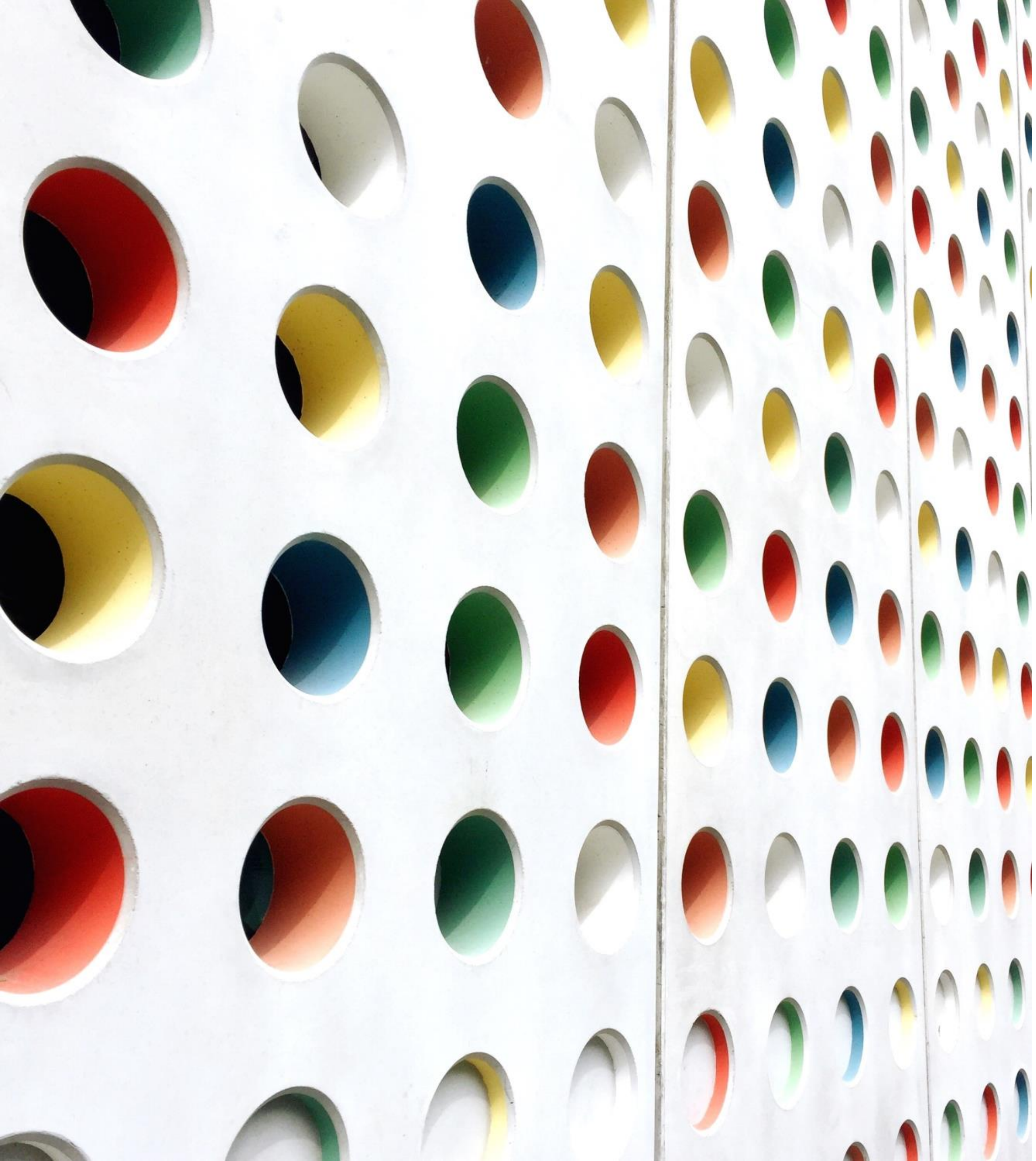


CREDIT-ONE

C2T1- DATA SCIENCE FRAMEWORK REPORT

SUGITHA DEVARAJAN

PROBLEM

Over the past year or so **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. The bottom line is they need a much better way to understand how much credit to allow someone to use or, at the very least, if someone should be approved or not losing business if the problem is not solved right away.

GOAL

DEFINE THE PROBLEM WITHIN A DATA SCIENCE FRAMEWORK

IDENTIFY WHICH CUSTOMER ATTRIBUTES RELATE SIGNIFICANTLY TO CUSTOMER DEFAULT RATE

BUILD PREDICTIVE MODEL THAT CREDIT ONE CAN USE TO BETTER CLASSIFY POTENTIAL CUSTOMER ARE BEING AT RISK.

DATA

DATA is retrieved from MYSQL database.

From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. (hint: Artificial neural network is the only one that can accurately estimate the real probability of default.)

Attribute Information:

✓LIMIT_BAL: Amount of the given credit - it includes both the individual consumer credit and his/her family (supplementary) credit.

✓SEX: Gender (1 = male; 2 = female).

✓Education (1 = graduate school; 2 = university; 3 = high school; 0, 4, 5, 6 = others).

✓Marital status (1 = married; 2 = single; 3 = divorce; 0=others).

✓ Age (year).

✓PAY_0,PAY_2,PAY_3,PAY_4,PAY_5,PAY_6: History of past payment. We tracked the past monthly payment records (from April to September 2005) as follows: PAY_0 = the repayment status in September 2005; PAY_2 = the repayment status in August 2005 PAY_6 = the repayment status in April 2005.The measurement scale for the repayment status is: -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). BILL_AMT1 = amount of bill statement in September 2005; BILL_AMT2 = amount of bill statement in August 2005; . . .; BILL_AMT6 = amount of bill statement in April 2005

✓.PAY_AMT1 – PAY_AMT6: Amount of previous payment (NT dollar). PAY_AMT1 = amount paid in September 2005; PAY_AMT2 = amount paid in August 2005; . . .;PAY_AMT6 = amount paid in April 2005.

✓Default payment next month : client's behavior; Y=0 then not default, Y=1 then default

DATA SCIENCE FRAMEWORK - BADIR

BADIR – Business Question

Analysis Plan

Data Collection

Derive Insights

Recommendations

The reasons for proposing BADIR framework are

It helps the project to be more successful with stake holders being more involved and the analytics team to be more engaging and delivering the recommendations. BADIR is a detective approach to find the solution by starting the process with business question and then planning the analysis. Recipe based analysis for deriving insights and recommendations.

How will you manage the data for the project

- Data protection is the most valuable thing now and we as analyst take this issue seriously.
- We only work in the secure company network and only store our data in the secure locations behind firewalls.
- We take yearly security training, and our team are all certified security experts when it comes to data protection.
- We comply with company policies and we will make sure **Credit One** data is protected.

Any Known Issue

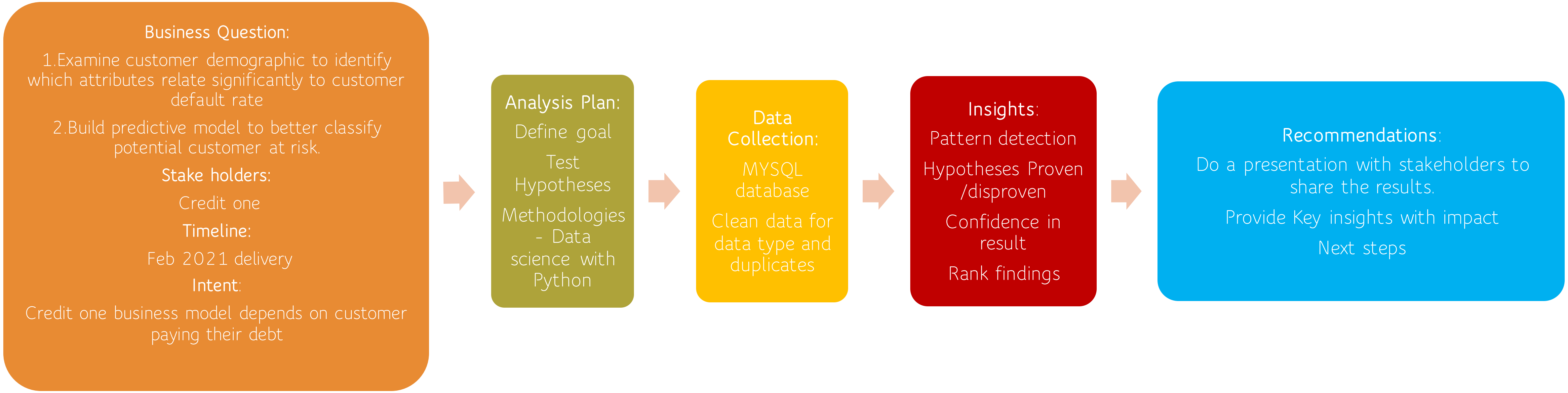
✓I found the header is some unknown column and the second row seems like the actual header. I am planning to drop the header and making the second row as the header for better analysis.

✓There were duplicates which will be dropped

✓Some of the columns were object type which were converted to integer for better analysis.

✓Trying to find correlation between different features to understand the relationship better to solve the business question.

Process Flowchart - BADIR



Business Question:

1.Examine customer demographic to identify which attributes relate significantly to customer default rate

2.Build predictive model to better classify potential customer at risk.

Stake holders:

Credit one

Timeline:

Feb 2021 delivery

Intent:

Credit one business model depends on customer paying their debt

Analysis Plan:

Define goal

Test Hypotheses

Methodologies - Data science with Python

Data Collection:

MYSQL database

Clean data for data type and duplicates

Insights:

Pattern detection

Hypotheses Proven /disproven

Confidence in result

Rank findings

Recommendations:

Do a presentation with stakeholders to share the results.

Provide Key insights with impact

Next steps

Initial Insights

Some of the questions I can answer looking at the data are

- What age group will likely to be defaulted?
- Will marital status play a role in a customer becoming defaulted?
- Did education feature contribute more to the customer predicting to be defaulted?
- Will female demographics be more likely to pay the debt or male?
- Will history of payment let us know if a customer will be defaulted?