C2 Project Scope and Preliminary Notes about the Data

Attribute Information (describes columns)

Features = 23 variables (try to simplify this done, if possible)

Target = default payment (1 = yes, 0 = no)

X1 = Credit amount (NT dollar) = individual consumer + family (supplementary) credit -- numerical continuous

X2 = Gender (1 = male, 2 = female) -- change for 0 and 1 -- categorical

X3 = Education level (1 = graduate school, 2 = university, 3 = highschool, 0,4,5,6 = others) -- what does "others" mean? -- categorical

X4 = Marital Status (1 = married, 2 = single, 3 = divorce, 0 = others) -- is others necessary? -- categorical

X5 = Age (year) - should be numerical discrete

Consider renaming X6 to X11 to "Payment History"

X6 through X11 refer to history of past payment -- should be numerical continuous

X6 = September 2005

X7 = August 2005

X8 = July 2005

X9 = June 2005

X10 = May 2005

X11 = April 2005

Considering changing columns -2 to 9 to "Repayment Status"

-2 = No consumption (what does that mean)

-1 = Paid in full

0 = revolving credit

1 - 8 (delayed payments by month, so 1 month up to 8 months) + 9 (9-month payment delay or more)

It appears that X18 to X23 relates directly to X6 to X11 -- maybe there's a way to combine them

X18 to X23 refer to previous payments (NT dollar) -- should be numeric continuous

X18 = payment for September 2005

X19 = payment for September 2005

X20 = payment for September 2005

X21 = payment for September 2005

X22 = payment for September 2005

X23 = payment for September 2005

Y = client's behavior (0 = not default, 1 = default) --- \* this is the target variable

Data Cleaning Process

1. Remove duplicate values
2. Replace missing data
   1. Replace with mean
   2. Replace with mode
   3. Remove missing data
   4. Do nothing
   5. Replace with zeros or some other constant data value (numeric, categorical, etc.)
   6. Others methods
3. Create clear names for each column
4. Determine if some columns of data can be removed from the dataframe:
   1. Remove duplicate data
   2. Simplifying columns (turning 4 columns of data into just 1 column
5. Determine the correct data type for each column of data
   1. Categorical discrete
   2. Categorical ordinal
   3. Numeric continuous
   4. Numeric discrete
   5. Time and date
6. Identify the key target variable
7. Make sure that categorical data makes sense and is encoded in a reasonable way
8. Check for skewed data distributions, as this could mean further manipulation on data
9. Check the min, mode, median and max values

Why does the DCP matter: Ensures understanding of the problem scope and fundamental questions related to the problem. Determines what edits need to be made to the dataset to improve the ability for the data science process to be carried out effectively, meaningfully and ethically.

Framework One - Zumel and Mount, Practical Data Science with R, chapter 1:

Define the goal The first step in a data science process is to define a measurable and quantifiable goal.

Why do the stakeholders want to do the project?

What do they need from it?

Why is their current solution inadequate?

What resources do you need?

How will the result of your project be deployed?

Collect and manage data This step includes identifying the data you need, then exploring and conditioning it. This is often the most time consuming step.

What data is available?

Will it help to solve the problem? Is it enough?

Is the data quality good enough?

Build the model Here is where you try to extract useful insights from the data in order to achieve your goals.

Which techniques might I apply to build the model?

How many techniques should I apply?

Evaluate and critique the model Once you have derived a model, you need to determine whether it meets your goals. If not, it’s time to loop back to the modeling step.

Is the model accurate enough to meet the stakeholders’ needs?

Does it perform better than "the obvious guess" and any techniques being used currently?

Do the results of the model make sense in the context of the real-world problem domain?

Present results and document Once you have a model that meets your criteria, you will present your results to your project sponsor and other stakeholders.

How should stakeholders interpret the model?

How confident should they be in its predictions?

When should they potentially overrule the model’s predictions?

Deploy and maintain the model Finally the model is put into production, but you still need to ensure that the model will run smoothly. In many cases this requires enhancement of the requirements based on customer feedback or in some cases fixing bugs.

How is the model to be handed off to "production"?

How often, and under which circumstances, should the model be revised?

Framework Two - BADIR (Jain and Sharma, Behind Every Good Decision, chapter 4):

Business question

What is the stated business question?

What is the intent underlying the question (e.g., what is the context, what is the impacted segment, and what are stakeholders’ current thoughts about the underlying reasons?

What business considerations (e.g., stakeholders, timeline, and cost) are likely to impact the analysis?

Analysis plan

What is the analysis goal?

What hypotheses are to be tested?

What data is required/available to test the hypotheses?

What methodology(-ies) will you employ?

What is the project plan (timeline and milestones, risks, phasing, prioritization, …)?

Data collection

From where can the data be obtained?

How must the data be cleansed and validated?

Insights

What patterns do you see in the data?

Are each of the hypotheses proven or disproven?

How much confidence should stakeholders place in the results?

How do you rank your findings in terms of quantified impact on the business?

Recommendation

How can you most effectively present the results of your analysis to your stakeholders (in terms they can understand and in alignment with information they’ll value)?

Note: A generic template for a recommendation presentation or report might include:

Objective

Background (optional)

Scope (optional)

Approach (optional)

Recommendations

Key insights with impact

Next steps