**Business question**

* **What is the stated business question?**

Why is there an increase in defaults?

* **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?**

An increase in defaults is a business problem for us, and it is our responsibility to be scoring the loans effectively. If we do not solve the problem soon, we could lose business and/or take losses.

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

A short timeframe is the main business consideration. As this is a business, cost is always a consideration.

**Analysis plan**

* **What is the analysis goal?**

The analysis goal is to find an empirical solution; i.e. a data based criteria explaining the increase in defaults that can be used to ameliorate the current situation.

* **What hypotheses are to be tested?**

No particular hypotheses, per se, just that the data can adequately explain the increase in defaults.

* **What data is required/available to test the hypotheses?**

We have historical data of 30,000 observations of 24 variables.

* **What methodology(-ies) will you employ?**

We will employ a variety of machine learning algorithms to see if we can come up with an accurate predictive model. Some data munging may be done, especially as relates to data exceeding the given definitions, as well as discretizing the nominal variables.

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

Timeline: ASAP

Milestones: Initial data review, predictive models

Risks: Unable to complete fast enough, models not accurate enough

**Data collection**

* **From where can the data be obtained?**

Historical data has been provided.

* **How must the data be cleansed and validated?**

Data is assumed to be valid; no outside validation of the data is needed nor feasible. Some variables exceed the definitions provided; the learning models should be able to incorporate these extra definitions, but we will not be able to explain as easily how or why they are impacting the learning models. Some data may need to be discretized, and perhaps other variables will be created/munged from the existing data.

Specifically:

**Education** is supposed to be a 1-3 ordinal scale, with 4 being ‘other’. 0, 5, and 6 show up in the data. Do these extra values maintain ordinality; i.e. is 0 post graduate and 5 and 6 < high school? Is one of these extra data elements a ‘not given’?

**Marriage** is supposed to be binary 1-2, or 3 = ‘other’. What does ‘other’ mean? 4 shows up in the data, is that a ‘not given’ or does it have other meaning?

**Repayment status** is supposed to be -1 for current, and 1-9 for ordinal lateness. The data shows up as -2 through 8. Did the data get shifted down by 1? Or did ordinal data get shifted down by 1, -1 is correct, and -2 means some prepayment?

**Bill amounts** show some negatives. Are these accidental overpayments or errors? Or does this list include loans to the company from others?

Other variables (payment amounts, age, gender, and credit limit) all look fine.

**Insights**

This will be applicable once we get to Task 2+

* **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**

This will be applicable once we get to Task 2+

* **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**