# BUAN 4310 PROJECT 1

# Overview

In this project, we will form groups of 2-4[[1]](#footnote-1) students each. Each group is similar to a consultancy and will be engaged to solve various problems that require our expertise in data mining. For this project, each group is engaged to solve 2 problems: Only 2 there will be, Problem 1 and Problem 2. These are business problems that will be transformed to analytics problems with analytics-driven solutions.

**Please note the following comparisons between Problem 1 and Problem 2**

|  |  |
| --- | --- |
| **Problem 1** | **Problem 2** |
| Larger dataset | Smaller dataset |
| More straightforward | Less straightforward |
| Deliverable: Annotated presentation slides or video presentation | Deliverable: Detailed report |
| Same topic for all, but different slices of a much larger dataset | Requires topic selection from a list |

**VERY IMPORTANT**: As with all projects and assignments, please ensure that the work and analyses are solely the group’s and not taken from another source. Using someone else’s or another party’s work and analyses amounts to plagiarism and will be handled according to SU’s policies. While we can use various sources as guides or references, please kindly understand that for purposes of the course, it is important to demonstrate our ability to solve the problems using data mining techniques. In other words, we have to showcase our abilities.

**ALSO VERY IMPORTANT**: Since we will select our own group mates, intra group issues, such as workload distribution, commitment, disputes, etc, should be handled by the group members internally. It’s a 4000-level course, and we should be ready to work even with team members that we hardly know and make it work somehow. The jedi master is not likely to intervene in the event of an intra group dispute.

Please take this as a learning experience that perhaps forces us out of our comfort zones to adapt to our environment and teams. Nonetheless, since this is a class setting, there is a peer evaluation of each group member’s contributions towards the end of the quarter. Each group member will rate the other group members on a scale of 1 to 5. The average score will determine the percentage of each member’s total project score.

In addition, all deliverables should be professionally prepared. This is included implicitly in the requirements (especially since we are business students). While deliverables are not graded for grammatical and typographical errors (this is not an English or Literature course), the language should be professional, coherent, and comprehensible. If need be, please feel free to utilise the writing centre (please refer to the syllabus) prior to submission.

# 1 Problem 1 (22 points)

**Due: Nov 1, 2020**

After Steve Rogers replaced the Infinity Stones, Stark Enterprises has branched into the financial industry. Perhaps Steve Rogers changed something when he travelled back in time[[2]](#footnote-2). Since Mr. Stark is on a different timeline, they are short of analytical power. They would like to determine which customers are high-risk. The documentation is provided on Canvas.

Each group will be assigned a random sample of the larger dataset (about 550,000 records and 27 variables). From the dataset, loan (loan\_1.csv, loan\_2.csv, loan\_3.csv, … loan\_n.csv), we will build an appropriate model (based on at least 2 models) for the company and present the findings.

## 1.1 Deliverable

Each group will present the findings from Problem 1. This is a professional, formal presentation[[3]](#footnote-3) to be submitted as a video recording or a deck of annotated presentation slides, along with the slide deck or html markdown with explanation[[4]](#footnote-4). The annotations should include the presentation script as though the speaker(s) presented the project. These are to be included in the notes section of each slide. Not every member of the group needs to talk. It is ok to designate 1 or 2 of the best presenters to present. However, it is expected that the group split the workload evenly.

Even though the data are fairly clean, some degree of data cleansing is expected. Students should spend some time getting the data ready for analysis. If needed, merging with other data or using other data to supplement the analysis will be good, although not necessary.

Some level of domain knowledge may be required to know how to explore and analyse the data. As consultants to your client, it is necessary to understand the business domain so as to make appropriate recommendations during the presentation.

Be sure to document your training-validation split and explain your split ratio.

A list of references cited should be included at the end of the presentation as additional slides. In-text citations are not necessary.

The presentation should include the following at the very least:

* Problem
* Objective
* Describe the purpose of the data from the topic you selected (i.e., why was this data collected in the first place?).
* **Results of data exploration**
  + Describe the meaning and type of data (scale, values, etc.) for each variable in the data file.
  + Verify data quality: Are there missing values? Outliers? Are those mistakes? How do you propose to deal with these problems (if any)?
  + Basic descriptive analyses on the data
    - Summary statistics and distributions for the most important attributes and describe what they mean or if you found something
    - Should existing variables be transformed? Why?
  + Various visualisations (at least 6 interesting visualisations). Provide an interpretation for each visualisation. Explain for each attribute why you chose the visualisation, what story does it tell, why is it important, etc.
    - Explore relationships between attributes: Look at the attributes via scatter plots, correlation, cross-tabulation, group-wise averages, etc. as appropriate.
* Identify and explain interesting relationships between variables and the outcome you are trying to predict (or cluster).
* Explain how you define and measure the outcome variable from the dataset.
* At least 2 models and their corresponding results and diagnostics[[5]](#footnote-5)
  + Be sure to include the technical details of the analysis on separate slides. These include data transformations, additional data (if any), criteria for variable selections, model specifications, etc. Normally, for a business audience, we may skip some of these. But if questioned by a more technical audience, these technical details may come in handy.
* Results of final model and rationale for selection **(complete together)**
  + Justify why you selected the final model from those you built
* Recommendations to your client (some research on the problem domain will be needed here)
* Sustainability of the project. Is it practical to collect these data in the long run (business consideration)?

Are there other variables that could be added to the data or created from existing ones? Which ones?

Next steps?

## 1.2 Problem 1 Rubric

The grading rubric is given below[[6]](#footnote-6). Approximately 10% of the grade is allocated to truly exceptional work. This involves going well beyond the expectations, thinking well outside the box, putting in very substantial effort, and providing additional explorations and manipulations[[7]](#footnote-7).

The grading rubric[[8]](#footnote-8) is given below. Approximately 10% of the grade is allocated to truly exceptional work. This involves going well beyond the expectations, thinking well outside the box, putting in very substantial effort, and providing additional explorations and manipulations[[9]](#footnote-9).

|  |  |  |  |
| --- | --- | --- | --- |
| **Data visualisations**  **5 points** | 5 points  Used very accurate and suitable visualisations on well-manipulated data. Professional, clear and well-labelled visualisations that are supported by in-depth explanations/ discussions. Provides clear and interesting relationships among the variables that are communicated effectively to business people. | 2.5 points  More data manipulation is needed to arrive at more effective visualisations. Fairly clear and labelled visualisations that are somewhat supported by fair explanations/ discussions. Provides some clear and some interesting relationships among the variables that may or may not be communicated effectively to business people. | 0 points  Poorly labelled and constructed visualisations that do not provide insights to the problem, for eg, mis-labelled charts, unclear legends, over-complex charts that do not convey messages easily to the layperson or business person, etc. |
| **Data understanding**  **5 points** | 5 points  Fully demonstrates critical understanding of the data, including the data types, missing data, outliers, distributions, and descriptives. Clear explanations of what this understanding means to business people and what they can be used for. Provides interesting insights that are understandable and communicated effectively to business people. | 2.5 points  Fairly demonstrates critical understanding of the data, including the data types, missing data, outliers, distributions, and descriptives. Fair explanations of what this understanding means to business people and what they can be used for. Provides somewhat interesting insights that may or may not be understandable and communicated effectively to business people. Some questions remain that are unclear about the data. | 0 points  Poorly demonstrates critical understanding of the data, including the data types, missing data, outliers, distributions, and descriptives. Poor explanations of what this understanding means to business people and what they can be used for. Provides little interesting insights that are may or may not be quite understandable and communicated effectively to business people. Many questions remain that are unclear about the data. |
| **Business understanding**  **5 points** | 5 points  Strong In-depth discussion of the business problem. Good description of the data, purpose of the data. Clearly explained to business people. | 2.5 points  Fair discussion of the business problem. Fair description of the data, purpose of the data. Some purposes of the data are not clearly explained to business people. | 0 points  Poor discussion of the business problem. Not clearly explained to business people. Some questions remain about why is there a need for these data and/or what they can be used for. |
| **Presentation quality**  **4 points** | 4 points  Professional and cohesive presentation. Covers key aspects and presents clear ideas to the audience. Excellent use of presentation visuals. | 2 points  Somewhat professional and/or somewhat non-cohesive presentation. Covers some key aspects and presents clear ideas to the audience. Fair use of presentation visuals. | 0 points  Unprofessional and/or non-cohesive presentation. Does not quite cover key aspects and present clear ideas to the audience. Weak use of presentation visuals. |
| **Exceptional work**  **3 points** | 3 points  Satisfied all other criteria. Well above and beyond the expectations of the assignment. Highly insightful recommendations that are well communicated to a business audience. Includes very sound additional analyses to support the findings, critical questions to ask of the data, and in-depth insights on analytical methods. Incorporates extensive, well-researched domain knowledge to support the analyses. Demonstrates critical understanding, such as strengths and limitations, of the appropriate methods based on the data exploration and extensive, accurate domain knowledge, to use for analysis. | 1.5 points  Satisfied all other criteria. A little beyond the expectations of the assignment. Somewhat insightful recommendations that may or may not be well communicated to a business audience. Includes some additional analyses to support the findings, critical questions to ask of the data, and in-depth insights on analytical methods. Incorporates some domain knowledge to support the analyses. Demonstrates fair understanding, such as strengths and limitations, of the appropriate methods based on the data exploration and domain knowledge, to use for analysis. | 0 points  Did not go above and beyond expectations. |

# 

1. No more than 4. However, if you wish to work in a pair, please have a good justification. For purposes of the project, which is a lot of work, solo undertakings are strongly discouraged. Furthermore, a big part of this field is to learn to work in teams. Often, we don’t even get to choose who we work with. [↑](#footnote-ref-1)
2. The problems with time travel (in general and in this case) are beyond the scope of the project (and course). That’s a conversation over a cup of tea! [↑](#footnote-ref-2)
3. Not everyone in the team needs to stand up and talk (stand up and shooooooooout!). It’s a collaborative effort, but each team may choose to delegate the presentation responsibility to 1 person. [↑](#footnote-ref-3)
4. In other words, we can submit either (1) video presentation recording + slides, (2) video presentation recording + html markdown, (3) annotated slide deck + slides, or (4) annotated slide deck + html markdown. Please let me know if you have questions. [↑](#footnote-ref-4)
5. The accuracy is likely to be very nice, although not perfect. So if your accuracy is just so-so, you may want to double check your work. Note that there are different ways to assess the quality of your solution. [↑](#footnote-ref-5)
6. Please note that in all written assignments, there will be an element of subjectivity. All grade disputes (if any) should be followed up within a week of releasing the corresponding grade for the graded deliverable. [↑](#footnote-ref-6)
7. One idea is to perform various data transformations in preparation for proposed data analysis methods and to arrive at very interesting results. This naturally, takes a lot of time. [↑](#footnote-ref-7)
8. One idea is to perform various data transformations in preparation for proposed data analysis methods and to arrive at very interesting results. This naturally, takes a lot of time. [↑](#footnote-ref-8)
9. One idea is to perform various data transformations in preparation for proposed data analysis methods and to arrive at very interesting results. This naturally, takes a lot of time. [↑](#footnote-ref-9)