BUAN 4310 Project

Albers School of Business and Economics

Warning: LOOOOOOOOOOONG document 😊

**Table of Contents**

[Overview 2](#_Toc53655770)

[1 Problem 1 (22 points) 3](#_Toc53655771)

[1.1 Deliverable 3](#_Toc53655772)

[1.2 Problem 1 Rubric 5](#_Toc53655773)

[2 Problem 2 (26 points) 7](#_Toc53655774)

[2.1 Topics 7](#_Toc53655775)

[2.1.1 Fantastic Houses and Where to Find Them 8](#_Toc53655776)

[2.1.2 With Great Power Comes Great Responsibility 8](#_Toc53655777)

[2.1.3 Nebula on Earth 8](#_Toc53655778)

[2.1.4 Carrie 8](#_Toc53655779)

[2.1.5 From Diagon Alley to the Muggle World 9](#_Toc53655780)

[2.1.6 Lemons 9](#_Toc53655781)

[2.1.7 Gringotts 9](#_Toc53655782)

[1.1.8 Dumbledore’s Additional Condition 10](#_Toc53655783)

[2.1.9 A Visit from Akiridion-5 11](#_Toc53655784)

[2.1.10 R&D From Trollmarket 11](#_Toc53655785)

[2.2 Deliverable 12](#_Toc53655786)

[2.3 Problem 2 Rubric 13](#_Toc53655787)

[Appendix Project Strategies 16](#_Toc53655788)

[Problem 1 16](#_Toc53655789)

[Presentation (annotated slides or video) 16](#_Toc53655790)

[Problem 2 16](#_Toc53655791)

[Report of your analysis 16](#_Toc53655792)

# 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. (Jay)
  + Verify data quality: Are there missing values? Outliers? Are those mistakes? How do you propose to deal with these problems (if any)? (Jay)
  + 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
  + 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. |

# 2 Problem 2 (26 points)

**Due: Nov 15, 2020**

Each group will **SELECT ONLY ONE** of the following topics. The descriptions are given in the corresponding documentation. Doing a different topic or changing topics halfway will require the group to re-do some things and cause unnecessary delays. Hence, it is very important to explore the data first before deciding. In some cases, the resultant models may not be that good. In reality, we don’t always get nice models. Often, we have to do more data preparation to get very good models[[10]](#footnote-10). Furthermore, sometimes, we may only be able to get a good model from a subset of the data. In other words, the model may only apply to certain groups of records. This naturally, requires exploration (i.e. a lot of work!). Given our time constraint, if we have sufficiently exhausted all our resources by the deadline, report the best model and note the results and diagnostics of the other models built[[11]](#footnote-11) to justify your selection. In this case, make appropriate recommendations to your client. Be professional!

## Topics

Most of the given topics are rather unstructured, and require considerable thinking and exploration. In other words, they’re tedious and time consuming (and challenging?). Any group that is more comfortable with a more structured Problem 1 can attempt Topic 1.1.7 instead.

Where unspecified, split the data into training and validation sets. Be sure to document and explain your split ratio.

### 2.1.1 Fantastic Houses and Where to Find Them

With escalating home prices in King County, an aspiring NoMaj (i.e. muggle), Jacob Kawalski, who has been sleepless because Queenie decided to join the dark side, launch a real estate business. However, he needs to understand the real estate market.

The dataset (kc\_house\_data\_2.csv) includes home prices in King County. The corresponding documentation (Real Estate in King County documentation.docx) provides the variable definitions. Using various data mining techniques, explore the data and build a suitable model to help Jacob determine what predicts home prices in King County.

### 2.1.2 With Great Power Comes Great Responsibility

Peter Parker has had a very busy year. Unlike many students, he has an unmanageable amount of schoolwork. In addition, he has to save the world as his alter ego. One day, he decided to take a break from his responsibility and outsource his responsibility to Jessica Jones during his coming summer vacation. He wants to start planning. Since summer is the time to stay indoors and hide from the sun[[12]](#footnote-12), he will get video games to play and trade. Obviously, he’s broke because he spent his extra time as his alter ego instead of working part time.

The dataset (vgsales.csv) includes video game sales. The corresponding documentation (Video\_game\_sales documentation.docx) provides the variable definitions. Using various data mining techniques, build a suitable model to help Peter determine what predicts video game sales in different regions.

### 2.1.3 Nebula on Earth[[13]](#footnote-13)

Anyone who is not tired of bank problems can try this (Problem 2 is a bank problem). This may also be better for those who are Finance or Accounting students. Nebula decided to stay on Earth after the snap was undone. She found a job at a bank with her data mining skills. However, she does not have access to her memories and cannot perform as well as expected. She sought out a consulting group and found you.

The bank wants to develop a stronger marketing campaign. To do so, they need to identify what determines a customer’s decision whether or not to take up a term deposit. Nebula was given this task. Using various data mining techniques on the given data (bank-full.csv), help her develop a suitable model. The corresponding documentation (Bank\_data\_documentation.docx) provides the variable definitions.

### 2.1.4 Carrie[[14]](#footnote-14)

“When 69 meets 40, there’s a single stop light town.”

Carrie Underwood, from the wonderful state of Oklahoma, wants to do something for her hometown. Oklahoma releases data on state employee purchases using the p-card. She wants to learn more about the purchase patterns of state employees to see what can be improved to reduce unnecessary spending. Carrie wants to select one very lucky group to work with. So she will compare the data mining submissions from various groups (using historical data) before she decides whom she wants to work with.

Using the data, find interesting purchase patterns and build a suitable model to classify 2015 transactions as necessary or unnecessary. There are 4 years of data (p\_card\_2012.csv, p\_card\_2013.csv, p\_card\_2014.csv, p\_card\_2015.csv) and 2015’s data only covers July – Dec. Please note that the “unnecessary” variable is entirely hypothetical for all years. The corresponding documentation (OK\_Purchase\_Data\_Documentation.docx) provides the variable definitions.

### 2.1.5 From Diagon Alley to the Muggle World

Weasley’s Wizard Wheezes are doing very well in Diagon Alley. The founders Fred and George have diversified their investment to sell regular products in the muggle world. However, things aren’t quite the same there. They have to use data mining to predict sales in their supermarket. So they sought your expertise.

Using the data, build a suitable model using the training data set (Train.csv) that can predict total sales in the validation set (validation.csv). Note that the data are already separated into training and validation sets. The corresponding documentation (Supermarket\_sales\_documentation.docx) provides the variable definitions.

### 2.1.6 Lemons

Lightning McQueen has retired. He is working with Mater to help identify lemons. Their new venture was inspired by their adventure in Europe. Lemons often have issues that mechanics cannot address, making them very costly if acquired.

Using the training data on used car sales (training\_4.csv), build a model to identify which cars have a higher chance of being a lemon. This will help dealerships acquire the best inventory possible. The model will predict if the cars in the validation set (validation\_4.csv) purchased at the auction is a lemon. The corresponding documentation (Lemons\_documentation.docx) provides the variable definitions.

### 2.1.7 Gringotts

If we really, really like banks, Gringotts Wizarding Bank is branching out into the muggle world[[15]](#footnote-15), just like George and Fred. They have begun a programme to encourage its existing muggle customers to borrow via a consumer loan program. The bank has promoted the loan to 5000 customers, of whom 480 accepted the over. The data are available in the file UniversalBank.csv. The bank now wants to develop a model to predict which customers have the greatest probability of accepting the loan, to reduce promotion costs and send the offer only to a subset of its customers. And hence, they have come to you.

* Using various data visualisation techniques and descriptive statistics, explore the data and report interesting findings. If nothing is interesting, report accordingly[[16]](#footnote-16).
* Build a kNN, and a classification tree using Personal Loan as the target variable.
* For each model, create a data frame showing the actual outcome and predicted outcome. For simplicity, only show the first 10 rows.
* Report the confusion matrices and compare the models.
* Advice your client accordingly.
* Use your models to predict the outcome for the following customers.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Customer | Customer 1 | Customer 2 | Customer 3 | Customer 4 |
| Age | 40 | 25 | 59 | 32 |
| Experience | 10 | 6 | 30 | 5 |
| Income | 84 | 50 | 120 | 60 |
| Family | 2 | 1 | 3 | 1 |
| CCAvg | 2 | 1.8 | 1.9 | 3.5 |
| Education | 2 | 1 | 3 | 2 |
| Mortgage | 0 | 120 | 0 | 250 |
| Securities Account | 0 | 0 | 0 | 1 |
| CD Account | 0 | 0 | 1 | 1 |
| Online | 1 | 1 | 1 | 1 |
| Creditcard | 1 | 1 | 0 | 1 |

### 1.1.8 Dumbledore’s Additional Condition

This might appeal to those interested in the healthcare industry. In Snape’s attempt to find an antidote for Dumbledore’s infection, he learnt that Dumbledore has diabetes. Since this is a muggle condition, Snape is seeking your help to learn more about it. Given the time constraints of the term, Snape has asked you build a suitable model to address any one of the following:

* if a patient will be readmitted within a 30 day period or not,
* what the total number of days a patient will spend in the hospital, given their history and specifics of the encounter.

Each record in the given dataset[[17]](#footnote-17) (diabetic\_data.csv) comprises a patient record. The documentation (documentation\_diabetes.docx) provides the variable definitions.

### 2.1.9 A Visit from Akiridion-5

Aja and Krel of House Tarron on Akiridion-5 are stuck on Earth. Knowing little about our pale blue dot gloriously named after dirt and the challenges living here, Señor Uhl of Arcadia Oaks High School has asked you to share with them your investigation on the relationships between county demographics (such as socioeconomic inequality), pharmaceutical interventions and non-pharmaceutical interventions, on the infections and mortality of the coronavirus. House Tarron is interested to learn about which counties have better or worse experiences with the coronavirus and how to potentially reduce infections and mortality from the coronavirus by investigating what is working or not working at the county level (i.e. what are the predictors, and to what extent).

This dataset comprises several variables and each record corresponds to a US county. The documentation (variable\_definition\_v4.xlsx) provides the variable definitions.

### 2.1.10 R&D From Trollmarket

This may appeal to finance majors. This topic is better suited to those who are familiar with various financial ratios.

With victory over the Gum Gums, the trolls from Trollmarket are learning about the human world. One thing they find confusing is R&D. R&D is sometimes regarded as non-essential for businesses. However, innovation is tied closely to R&D, which in turn requires investments (i.e. R&D expenditure). Using the data provided, they want to investigate what financial characteristics drive R&D expenditure, so as to know what are key factors that can predict a company’s decision to invest in R&D. Each record in the dataset corresponds to a company-year record (i.e. a company in a specific year).

The data were obtained from WRDS from a combination of 2 datasets. A subset of the entire dataset (fin\_performance\_rq\_subset1.csv) will be used for this investigation. Note that the documentation may not be very clear or complete for this one (as with many real world problems). For purposes of the class, please feel free to drop undocumented variables. The documentation comes in 2 files: fin\_ratio.pdf and additional\_documentation\_fin\_rq.docx.

## 2.2 Deliverable

For Problem 1, we are to submit a professionally prepared report with the details given above. Submit the report on Canvas by the due date. Until master yeoda figures out how to do a group submission, each member of the group should make the same submission.

The report should at the very least include the following elements:

1. Problem definition
   * Here, we formally define the problem
   * Why is this necessary?
2. Objective
   * Formal objective
   * What does this achieve for Stark enterprises?
3. Descriptive analysis of the data
   * Use various data exploration techniques to explore the data
   * These include various charts (at least 5 useful visualisations), aggregates, and descriptive statistics
   * Discuss the findings
     1. Explore relationships between attributes: Look at the attributes via scatter plots, correlation, cross-tabulation, group-wise averages, etc. as appropriate.
     2. Provide an interpretation for each chart. Explain for each attribute why you chose the used visualisation.
4. Final data mining model[[18]](#footnote-18)
   * This is final model selected to solve the problem. Note that to arrive at a final model, there must have been at least 2 models for selection.
   * Explain the algorithm and how it works (for a business audience)
   * Use charts as needed to visualise the model (for example, you can include the decision tree if you decided to use a classification tree)
   * Explain the inclusion/exclusion of any variable(s)
   * Assumptions
5. Interpretation of findings
   * Explain the findings to a business audience
   * What are the interesting variables? How so?
   * How will this help Stark Enterprises?
   * What is/are interesting?
6. Quality of the model
   * Discuss how good is the final model
   * How can this be used?
7. Recommendations
   * Based on the findings, what are the recommendations? (Some research on the problem domain will be needed here)
8. Sustainability of the project. Is it practical to collect these data in the long run (business consideration)? How often should this model be updated?
9. What other data can be used to enhance the model?
10. As with all published works, referencing is important. All facts/concepts must be attributed. However, there is no need for a formal citation style. As long as it is a complete citation (i.e. website title, URL and date if available; not just URL). The use of footnotes or endnotes would be good (and easy to do), as APA style tends to make it hard to read for a business audience.
11. Appendix
    * This is where we elaborate on the more technical aspects for academic purposes[[19]](#footnote-19), or for a selected business audience that is interested in learning more.
    * Include the following
      1. The techniques applied, and how.
      2. Data preparation. Explain the process and why you did what you did for all techniques applied
      3. Results of each technique applied and/or each model built
      4. Why the final data mining model was chosen?
      5. Strengths and limitations of different models
      6. How useful were each of these models developed? Why?
      7. What other data could be useful?
      8. Additional details as needed

## 2.3 Problem 2 Rubric

The grading rubric[[20]](#footnote-20) 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, manipulations and additional models for evaluation[[21]](#footnote-21).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Outstanding | Acceptable | Troll |
| **Addressing the requirements** | 4 points  Included the required components of the assignment. Used relevant and quality sources to support ideas. Properly formatted bibliography. | 2 points  Included some of the required components of the assignment. Used relevant and average quality sources to support ideas. Weakly properly formatted bibliography. | 0 points  Did not include most of the required components of the assignment. Did not use references. |
| **Business understanding** | 4 points  Strong In-depth description of the business problem. Good description of the data, purpose of the data. Clearly explained to business people. Very strong recommendations based on the findings that can be applied to business strategies. | 2 points  Fair description 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. Fair recommendations based on the findings that may or may not be applicable to business strategies. | 0 points  Poor description 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. Poor recommendations based on the findings that may or may not be applicable to business strategies. |
| **Data understanding** | 4 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 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. |
| **Data modelling and evaluation** | 7 points  In-depth explanations for all the variables used and why they are appropriate for analysis. Very well justified model. Strong selection criteria applied when selecting the final model. Very clearly identified and explained the findings and how the model can be used. | 3.5 points  Fair explanations for all the variables used and why they are appropriate for analysis. Fairly justified model. Fair selection criteria applied when selecting the final model. Fairly identified and explained the findings and how the model can be used. | 0 points  Poor explanations for all the variables used and why they are appropriate for analysis. Poorly justified model. Poor selection criteria applied when selecting the final model. Poorly identified and explained the findings and how the model can be used. |
| **Visualisations** | 4 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 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. |
| **Exceptional work** | 3 points  Satisfied all other criteria. Well above and beyond the expectations of the assignment. Includes very well justified and executed additional analyses and models to support the findings, critical questions to ask of the data, and in-depth insights on analytical methods. Incorporates very well researched and sound domain knowledge to support the analyses. Demonstrates critical understanding, such as strengths and limitations, of the applied (and not applied) methods based on the data exploration/analysis and extensive, accurate domain knowledge, to use for analysis. | 1.5 points  Satisfied all other criteria. A little beyond the expectations of the assignment. Includes some additional analyses and model 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 applied (and not applied) methods based on the data exploration/analysis and domain knowledge, to use for analysis. | 0 points  Did not go above and beyond expectations. |

# Appendix Project Strategies

## Problem 1

### Presentation (annotated slides or video)

* Data exploration
* Use various data visualisations to tell your client how things are (descriptive analytics)
* Build predictive model
  + Categorical outcome variable
  + kNN, classification tree, logistic regression[[22]](#footnote-22)
* Unsupervised learning (as needed)
* It is possible that the model works for a subset of the data or works differently for different subsets.

## Problem 2

### Report of your analysis

* Data exploration
  + Use various data visualisations to tell your client how things are (descriptive analytics)
* What models did you build? (for predictive analytics)
* If you have a numerical outcome variable
  + No transformation
    - Regression tree, Linear regression
      * Identify predictors
  + Transform into categorical
    - kNN, classification tree, logistic regression[[23]](#footnote-23)
* If you have a categorical outcome variable
  + kNN, classification tree, logistic regression[[24]](#footnote-24)
* Unsupervised learning (as needed)
* It is possible that the model works for a subset of the data or works differently for different subsets.

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)
10. Data analysts’ jobs are mostly data preparation (maybe 75% - 80%), compared to the actual analysis. [↑](#footnote-ref-10)
11. Naturally, to say we have exhausted all our resources, we should have built several decent and well-justified models. These have to be fully explained in an Appendix. [↑](#footnote-ref-11)
12. Ok, this is master yeoda’s biased opinion. [↑](#footnote-ref-12)
13. Source: S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, In press, http://dx.doi.org/10.1016/j.dss.2014.03.001 [↑](#footnote-ref-13)
14. Source: Purchase Card Data. State of Oklahoma. https://data.ok.gov/tags/purchase-card [↑](#footnote-ref-14)
15. I am running out of ideas for scenarios. [↑](#footnote-ref-15)
16. Data exploration is unstructured by nature. So this part is still rather open and ambiguous. [↑](#footnote-ref-16)
17. Beata Strack, Jonathan P. DeShazo, Chris Gennings, Juan L. Olmo, Sebastian Ventura, Krzysztof J. Cios, and John N. Clore, “Impact of HbA1c Measurement on Hospital Readmission Rates: Analysis of 70,000 Clinical Database Patient Records,” BioMed Research International, vol. 2014, Article ID 781670, 11 pages, 2014. [↑](#footnote-ref-17)
18. Groups are expected to run several models to determine which is the best performing and most appropriate one(s). The details of these different runs are to be written in the Appendix. In an actual scenario, we probably will not have to include these technical methodological write-ups. But since this is a class setting, we should be able to explain why we did what we did. [↑](#footnote-ref-18)
19. For purposes of the project, omitted details on the data analysis can imply that the necessary or otherwise data pre-processing step was not taken satisfactorily. [↑](#footnote-ref-19)
20. 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-20)
21. One idea is to use clustering to arrive at suitable and useful clusters for data analysis, including various data transformations to arrive at very interesting results. This naturally, takes a lot of time. [↑](#footnote-ref-21)
22. Ibid. [↑](#footnote-ref-22)
23. We won’t have time to do logistic regression. So there’s no need to build a logistic regression model. [↑](#footnote-ref-23)
24. Ibid. [↑](#footnote-ref-24)