BUAN 4310 Project Problem 1

Albers School of Business and Economics

Warning: LOOOOOOOOOOONG document 😊

**Table of Contents**

[Overview 2](#_Toc146099484)

[Problem: Cars 2](#_Toc146099485)

[Directions 3](#_Toc146099486)

[Deliverable 4](#_Toc146099487)

[Project Problem 1 Rubric 4](#_Toc146099488)

# Overview

In this project, we will stay in our groups of 1-2[[1]](#footnote-1) students. Each group is similar to a consultancy and will be engaged to solve various problems that require our expertise in machine learning.

**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; the jedi mind trick has its limits ☺ 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 an optional 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. We may share this on our GitHub pages for our job hunt, so please ensure the deliverable is presentable. If need be, please feel free to utilise the writing centre (please refer to the syllabus) prior to submission.

# Problem: Cars

Lightning McQueen is lost. After Mack accidentally dropped him off in the middle of nowhere, he wondered to Radiator Springs where he found a new home. His new friend Mator wants to start a used car resale business but does not understand how to set prices for cars. Armed with an incomplete dataset about used car sales, they would like your expertise to predict whether a used car price should be high or low. Please note that the dataset does not have exact prices (and hence incomplete). Instead, they will use high vs low to determine how to set the price (presumably based on prevailing market rates) and market used cars. High priced ones will be marketed differently from low priced ones. While the price for some used cars are obvious, others are not. For eg, an old car without many features is not likely to fetch a high price. But if the old car has many features that are in demand, it may be able to fetch a high price.

Each group will be assigned a random sample of the larger dataset (about 30,000 records and about 60[[2]](#footnote-2) variables and 1 target variable to train and validate). From the given dataset (car\_train\_class\_1.csv, car\_train\_class\_2.csv, car\_train\_class\_3.csv, … car\_train\_class\_n.csv), we will build and select an appropriate model for the company and predict the risk (i.e. the target variable) of the new customers given (car\_test\_1.csv, car\_test\_2.csv, car\_test\_3.csv, ….., car\_test\_n.csv )[[3]](#footnote-3).

Refer to the given documentation (used\_cars\_description.xlsx) that describes the variables.

As with all business problems, there is a degree of ambiguity involved.

# Directions

For purposes of the project, the following is a broad outline of how to approach the problem. Please note that in practice, we may not be given these.

* Select the appropriate variables to train **at least 2 different models** to predict the target variable[[4]](#footnote-4).
* Validate the model.
  + This may require a bit of work as we are expected to select the appropriate predictors and adjust the model specifications to build the best model.
  + We also need to evaluate the models to know which is the best
* Select your best model (out of the >= 2)
* Explain your results and evaluate your best model.
* Use the model to predict the price level of the new used cars given.

Note: Given the nature of the problem, it may be difficult to get a fantastic model/solution. But we should be able to get a reasonable one that can be used for prediction. Credit will be given on the rigour of the solution, which may include data transformations (as needed).

# Deliverable

Each submission comprises an R script report or rmd report (in word or html) that shows your codes, output, your predictions, and your explanation. This can be part of your portfolio for future use, so please prepare it professionally.

Even though the data are fairly clean, some degree of data cleansing may be needed. Students should spend some time getting the data ready for analysis. If a different application is used for the cleansing[[5]](#footnote-5), please note that in your explanation.

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 try to understand the business domain so as to select the appropriate predictors in the models.

The single **word/html/pdf markdown** should include the following at the very least:

* Brief problem description (as given)
* Objective (as given)
* Describe the data (i.e., what is this dataset about? What are the key variables?). There’s no need to do an EDA here, unless it’s used to support/explain the results.
* Details and explanations on data transformations (if any)
* Your models
  + Explain your pre-processing, specifications, explanation of set up and variable selection, results, interpretation, model evaluation[[6]](#footnote-6).
  + Justify the selection of your best model
* Discuss how good your best model is
* Predictions for the new customers using your best model

Late submissions without an approved extension are subject to a late penalty of 5% per day late. Please let the jedi master know if you have difficulty with the deadline and we’ll try to work something out.

# Project Problem 1 Rubric

The grading rubric is given below[[7]](#footnote-7). Approximately 5% of the grade is allocated to truly OUTSTANDING work. This involves going well beyond the expectations, thinking well outside the box, putting in very substantial effort, and providing additional explorations and manipulations[[8]](#footnote-8).

|  |  |  |  |
| --- | --- | --- | --- |
|  | OUTSTANDING | ACCEPTABLE | TROLL |
| **Business understanding**  **2 points** | 2 points  Strong description of the business problem. Good description of the data, purpose of the data. Clearly explained to business people. | 1 point  Fair description of the business problem; and/or fair description of the data, purpose of the data; and/or 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 the data and/or what they can be used for. |
| **Data understanding**  **3 points** | 3 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. | 1.5 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. 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. Many questions remain that are unclear about the data. |
| **Data mining**  **15 points** | 15 points  Correct set up of the models. Good justification of the variables. Good explanation of the results. Good justification for the best model. Correct classification of the new records. | 9 points  Somewhat correct set up of the models; and/or fair justification of the variables; and/or fair explanation of the results; and/or fair justification of the best model; and/or fair classification of the new records. | 0 points  Irrelevant. |
| **Presentation quality**  **2 points** | 2 points  Professional and cohesive presentation. Covers key aspects and presents clear ideas to the audience. | 1 point  Somewhat professional and/or somewhat non-cohesive presentation. Covers some key aspects and presents clear ideas to the audience. | 0 points  Unprofessional and/or non-cohesive presentation. Does not quite cover key aspects and present clear ideas to the audience. |
| **Exceptional work**  **1 point** | 1 point  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 and extensive, accurate domain knowledge, to use for analysis. And/Or submitted on time. | 0.5 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 and domain knowledge, to use for analysis. And/Or submitted late | 0 points  Did not go above and beyond expectations. |

1. If need be, we can do 3. But it’ll be harder to coordinate among yourselves. [↑](#footnote-ref-1)
2. 😊 [↑](#footnote-ref-2)
3. Each group is given a different set of new customers. [↑](#footnote-ref-3)
4. Some degree of data exploration may be needed to understand the data and justify our variable selections. We need at least 2 models to be able to select the best one. These could be different classification algorithms and/or the same classification algorithm but with different justified and reasonable settings. [↑](#footnote-ref-4)
5. Such as removing missing values. [↑](#footnote-ref-5)
6. For a business audience, sometimes we may skip some of these. But if questioned by a more technical audience, these details may come in handy. [↑](#footnote-ref-6)
7. 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-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)