Midterm Project 2

Small Business Data Set Version
MTH 3270 Data Science
Due Mon., May 3

Rules

You must do your own work, and you're only allowed to speak about this project with the instructor (Grevstad).

All analyses (data wrangling, visualizations, statistical summaries, etc.) must be done using **R** (except by permission of the instructor).

The projects are **due** in **Canvas** as a **pdf** file no later than **Monday, May 3, 2021** at **11:59 PM**.

Instructions

The project will use the data set from the Sixth International Conference on Establishment Statistics (ICES VI) student contest focusing on the analysis/visualization of economic data from the 2007 Survey of Business Owners.

The **data set** and a **data dictionary** describing the variables in it are obtained via the links below. Save the **csv** file and read it into R using `read.csv()` (and don't forget `header = TRUE` and `stringsAsFactors = FALSE`). Check **Canvas Announcements** and/or your **email** regularly in case there are important announcements about this project.

The **data set** is here (or go to the third website below, click **Student Contest – Data Analysis and Visualization**, and look for the **.csv** data file):

`https://ww2.amstat.org/meetings/ices/2021/studentcontest/track2sbo.csv`

A **data dictionary** is here:

`https://ww2.amstat.org/meetings/ices/2021/pdfs/contestdata_DataDictionary.pdf`

More **information** about the data and student contest can be found here:

`https://ww2.amstat.org/meetings/ices/2021/studentcontest.cfm`

You *might* need to do some data wrangling and tidying (which *might* involve selecting columns, adding new columns, filtering rows, grouping by a categorical variable, etc.).

Tasks

Your **tasks** are:

1. Carry out a **multiple regression analysis**. You may choose any response variable (Y) for your model, but it must be a numerical variable (*not* categorical). Likewise, you may use any explanatory (X) variables, but they too must be numerical (*not* categorical). Note that a categorical variable that's been coded using integer values is still considered to be a *categorical* variable.

   - **Summarize** your fitted model by reporting the estimated model coefficients.
   - **Interpret** the estimated model (coefficients).
   - **Report** the value of at least one measure of **how well** the model **fits** the data (e.g. the R^2).

2. Carry out a **logistic regression analyses** for predicting whether a business is the **primary source of income** for the **first** owner based on other explanatory variables from the data set.

   For the response (Y) variable, you'll use the *dichotomous* PRMINC1 variable taking the value **1** if **yes** and **2** if **no**. You may use any explanatory (X) variable(s), but they must be numerical (*not* categorical).

   You should **recode** the PRMINC1 variable first, so it takes the value **1** if **yes** and **0** if **no**, for example (if your data set is named small_business) by typing:

   ```
   small_business <- mutate(small_business,
                            PRMINC1 = ifelse(PRMINC1 == 1,
                                             yes = 1,
                                             no = 0))
   ```

   to ensure the model estimates the probability of **yes** (*not* **no**).

   - **Summarize** your fitted model by reporting the estimated model coefficients.

3. Carry a *machine learning* **classification** procedure (decision tree, random forest, k nearest neighbor, or artificial neural network – your choice) for **predicting *one*** of the following **categorical** variables (your choice). You may use any explanatory (X) variables, but they must be numerical (*not* categorical).

→ **Whether** a business is the **primary source of income** for the **first** owner. You'll need to convert the **0** and **1** values (of the **recoded** `PRMINC1` variable) to `"character"` (so they won't be treated as numerical responses by the model-fitting function in R) using `as.character()` with `mutate()`.

→ **Education level** of the **first** business owner `EDUC1`. You'll need to convert the **1**, **2**, ..., **7** values to `"character"` (so they won't be treated as numerical responses by the model-fitting function in R) using `as.character()` with `mutate()`.

→ The **hours per week** spent managing or working the business by the **first** business owner `HOURS1`. You'll need to convert the **1**, **2**, ..., **7** values to `"character"` (so they won't be treated as numerical responses by the model-fitting function in R) using `as.character()` with `mutate()`.

Then

- **Summarize** your procedure: Indicate *which* **classification procedure** you used, *which* (categorical) **response variable** you were predicting, and *which* **explanatory variables** you used.
- **Report** the value of at least one measure of **how well** the model **predicts** (**classifies**) individuals, e.g. the *correct classification rate.*
- **Provide** an **example** of a **prediction** (**classification**) using your fitted classification model.

What to Turn In

1. A well-organized **write-up** as a **pdf** file (perhaps 3-7 pages) containing:

   (a) A **brief description** (e.g. 1-2 paragraphs) of any data wrangling and tidying you had to do in order to carry out tasks **1**, **2**, and **3** above.

   (b) Your **responses** addressing the **bullet items** under tasks **1**, **2**, and **3** above (*seven* bullet items total).

2. Your **R code** with **comments** (use `#`) indicating **what** each chunk of code does and **why** it does it, either as an **appendix** in your **write-up pdf** or as a separate **.R file** (as produced by RStudio's script editor).

Grading

Your **grade** will be based on:

1. Your attainment of **tasks 1-3** above.
2. Your **write-up**, including your **responses** addressing the seven **bullet items** (as described above).
3. The inclusion of and correctness of your **R code**.