Midterm Project 3

Small Business Data Set Version
MTH 3270 Data Science
Due Sat., May 15

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 **Saturday, May 15, 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 **two tasks** are:

1. Every machine learning procedure has at least one **tuning parameter**, whose value you choose, that controls the **model complexity**, that is, how closely the fitted model is able to conform to the data:

   → *Decision tree.* The tuning parameters are: 1) The **minimum size of a node** in order for a split to be attempted; 2) The **complexity parameter**, for which a split is only performed if it decreases the misclassification rate by this percent or more.

   → *Random forest*: The tuning parameter is the **number of variables** to use in each tree.

   → *K nearest neighbor*: The tuning parameter is the **number of neighbors**, ***k***.

   → *Artificial neural network*: The tuning parameter is the **number of hidden units**, ***k***.

   A poorly chosen tuning parameter value leads to **overfitting** or to **underfitting**. A good tuning parameter value does neither. In other words, a good tuning parameter value produces a fitted model that classifies or predicts **out of sample** observations well.

   Your **first task** is to separate the small business data set randomly into **75% training** and **25% testing** sets, then fit one of the above **machine learning** models (your choice) to the **training set** using using **at least three** different values of the **tuning parameter**, and compare the effectiveness of each model for **classifying** individuals in the **test set**. Your model should **predict *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 **values** of the **tuning parameter** you evaluated, and *which* **of these values** resulted in the *best* **classifier** of individuals in the **testing set**, e.g. which one had the highest *correct classification rate* for this set.

2. Your **second task** is to carry out a **cluster analysis** (*hierarchical* or *k means*, your choice) to group the businesses into ***k*** clusters, where ***k*** is in the range **2-5** (your choice). You must use **four or more** explanatory (X) variables in the cluster analysis, and they must be numerical (*not* categorical). It's your choice which ones to use.

   Then inspect whether the clusters seem to correspond to whether a business is the **primary source of income** for the **first** owner. To decide, look for whether businesses within clusters largely are or aren't the primary source for the first owner (use the variable `PRMINC1`). This can be an informal inspection or something more formal (e.g. computing a measure of ”purity” for each cluster) – your choice.

   (It's okay if the businesses *don't* cluster according to whether they're the primary source of income for the first owner.)

   You're allowed to use only a **subset** of **rows** (observations) because **clustering procedures** are memory hogs and are computationally intensive. For example, to use just businesses from Sector 54 (of the North American Industry Classification System) that are franchises, (if your data set is named `small_business`) you might type:

   ```
   small_business <- small_business %>%
     filter(SECTOR == 54 & FRANCHISE == 1) %>%
     select(EMPLOYMENT_NOISY, PAYROLL_NOISY, RECEIPTS_NOISY, PCT1:PCT4)
   ```

   (Above, `select()` is used to select just the numerical variables in the data set.)

   Then

   - **Summarize** your procedure: Indicate *which* **cluster analysis procedure** you used, *how many* **groups *k*** you used, and *which* **explanatory variables** you used, and *how many* observations ended up being in each of the k clusters (groups).
   - **Report** the results of your assessment of whether the clusters seem to correspond to businesses that largely are or aren't the **primary source of income** for the **first owner**.

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