CS x460
Practical Machine Learning with R
Final Challenge Problem
Due: March 28 at 6**:29 PM**
(submit using Canvas, no late submissions accepted, no exceptions)

Instructions

- Work independently
- No code sharing
- Send any questions have about the assignment directly to me via Canvas messaging or email

Submission

- **Submit on Canvas** / onlinelearning.berkeley.edu (no email submissions accepted, no exceptions)
- Create a reproducible .Rmd R Notebook, with your source code and answers to the questions below
    - **Assume the data files are located on the same directory as your .Rmd file**, it should knit on my machine with a single click
    - Include the problem number and some identifying text for each problem in your RMarkdown document along with your solutions
    - Be sure to include your name in the heading of your RMarkdown document
- Knit your RMarkdown to HML and submit it to Canvas as a file named:

    first_name_last_name_final_challenge_problem.html

- Also submit the RMarkdown (.Rmd file) you knitted as a file named:

    first_name_last_name_final_challenge_problem.Rmd

Example: allan_miller_final_challenge_problem.html

Problems

Problem 1

Use the dataset **birth_data.csv** having the following fields:

atRisk	Logical likely to need immediate emergency extra medical attention upon birth
PWGT	Numeric Mother's prepregnancy weight
UPREVIS	Numeric (integer) Number of prenatal medical visits
CIG_REC	Logical TRUE if smoker; FALSE otherwise
GESTREC3	Categorical Two categories: <37 weeks (premature) and >=37 weeks
DPLURAL	Categorical Birth plurality, three categories: single/twin/triplet+
ULD_MECO	Logical TRUE if moderate/heavy fecal staining of amniotic fluid
ULD_PRECIP	Logical TRUE for unusually short labor (< three hours)
ULD_BREECH	Logical TRUE for breech (pelvis first) birth position
URF_DIAB	Logical TRUE if mother is diabetic
URF_CHYPER	Logical TRUE if mother has chronic hypertension
URF_PHYPER	Logical TRUE if mother has pregnancy-related hypertension
URF_ECLAM	Logical TRUE if mother experienced eclampsia: pregnancy-related seizures

(posted on the Files/Final Challenge Problem) to build a model that determines whether a baby is *at risk*, i.e., needs immediate emergency care or extra medical attention immediately upon birth.

Requirements

- **Use caret for your solution**, see:

  https://topepo.github.io/caret/

- Use **caret-based data preparation** to prepare the data for your models.

- Use **caret to fit (1) logistic regression and (2) GBM models.**

- Use **caret to tune your models appropriately**.

- Use caret (and other methods, if desired) to **evaluate and compare the performance of each type of model using appropriate methods** (e.g., confusion matrix, ROC, AUC).

Be sure to document and explain the rationale and results from each step of the modeling process, explain and evaluate your results.

Problem 2

Use the dataset **ocdata.csv** (posted on Files/Final Challenge Problem) having the following fields:

education, income, women, prestige, census, type

to answer the questions below.

(Problem 2A) Fit a univariate OLSR (Ordinary Least Squares Regression) model, adhering to OLSR assumptions, predicting **income from prestige only**.

(Problem 2B) Fit a model of any type we discussed in class, using all meaningful predictors of income, to obtain the "best" results, using whatever method you wish.

Use of caret for Problem 2 may be helpful, but is not required.

Be sure to properly prepare your data, take steps to avoid overfitting, tune and evaluate the performance of your model, and provide a clear description and analysis of your results.

Caution: for your solution, *more is not better*. Use only an appropriate model and methods that provides value. Avoid unnecessary work.

Also, pay attention to the appearance and quality of your HTML product: attractive, crisp, and clear.

Remember, hand in only two files with your solutions on Canvas:

first_name_last_name_final_challenge_problem.Rmd
first_name_last_name_final_challenge_problem.html