fairchild_fake_data

Let's make some fake data

```
library(caret)
```

```
## Loading required package: lattice
```

```
## Loading required package: ggplot2
```

```
library(fakeR)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.6     v dplyr   1.0.4
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.1
## v purrr   0.3.4
```

```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## x purrr::lift()   masks caret::lift()
```

```
library(recipes)
```

```
##
## Attaching package: 'recipes'
```

```
## The following object is masked from 'package:stringr':
##
##     fixed
```

```
## The following object is masked from 'package:stats':
##
##     step
```

```
library(reticulate)
use_virtualenv("./venv", required = TRUE)
```

```
## Warning in data.class(current): NAs introduced by coercion
```

```
## Warning in data.class(current): NAs introduced by coercion
```

```
## [1] "Some unordered factors..."
## [1] "Numeric variables. No ordered factors..."
```

Preprocessing Step

1) Applying inclusion criteria, each obs must be complete.

- Fairness question here: what is lost by dropping all rows with an NA?

2) Mothers must be 18 or older

3) Maternal morbidity excluded
4) Alive babies with gestational age less than 21 weeks excluded
5) Multiple birth pregnancies excluded
6) Pregnancies that ended in fetal death due to external causes excluded.
7) Parity status of the mother was deducted from the number of prior births variable
8) A new class variable was annotated to specify the outcome of pregnancies more accurately.
9) Fetal death cases were divided into late and early stillbirth based on their gestational age, at 28 weeks.
10) Early stillbirth cases of less than 21 weeks were excluded because they are clinically defined as miscarriage cases.
11) Live births were divided into uncomplicated pregnancies, and PTB pregnancies ("pregnancies with delivery before 37 weeks of gestation").
    - We should get like 12,000,000 normal pregnancies, 1,000,000 PTB cases, 7924 early stillbirth cases and 8310 late stillbirth cases.
12) Continuous Variable mean-zero normalization and unit-variance normalization.
13) Nominal predictors were one-hot encoded
14) CDC data was partitioned into four sets; feature selection data, training data, validation data and test data along 10, 70, 10, 10 split.

```
# functions
normalize <- function(x)  {
  return ((x - mean(x)) / sd(x))
}
```

```
pregs_pp <- pregs %>%
  mutate(id = row_number()) %>%                      # adding an id
  relocate(id) %>%
  filter(if_all(where(is.numeric), ~ . >= 0)) %>%   # filter negative numeric values created by synthet
  drop_na() %>%                                      # 1
  filter(age >= 18) %>%                              # 2
  # 3 NOT DONE
  filter(g.weeks >= 21) %>%                          # 4 and 10
  #5 #6 #7 NOT DONE
  mutate(outcome = case_when(
    stillborn == 0 & g.weeks < 37  ~ "preterm",
    stillborn == 0 & g.weeks >= 37 ~ "normal",
    stillborn == 1 & g.weeks < 28  ~ "early stillbearth",
    stillborn == 1 & g.weeks >= 28 ~ "late stillbirth"
  )) %>%                                             # 8, 9, 11
  mutate(BMI = normalize(BMI)) %>%
  mutate(height = normalize(height)) %>%
  mutate(weight = normalize(weight)) %>%             # 12
  mutate(race.f = factor(race)) %>%
  mutate(education.f = factor(education)) %>%
  mutate(race = paste("race", race, sep = "_"),
         race_hot = 1,
         education = paste("education", education, sep = "_"),
         education_hot = 1) %>%
  spread(key = race, value = race_hot, fill = 0) %>%
  spread(key = education, value = education_hot, fill = 0) # 13 while preserving factors also
```

```r
any(is.na(pregs_pp))  # Should be False!
```

```
## [1] FALSE
```

```r
table(pregs_pp$race.f)  # Make sure the simulated data is complete enough
```

```
##
## amerindian      asian      black      white
##       2290       3113       2757       2698
```

```r
table(pregs_pp$education.f)
```

```
##
##       8th   College Doctorate         HS
##      2290      2757      3113      2698
```

Splitting data, and where to now?

> … caret (v. 6.0-82) was used for data partition [17].
>
> For conducting the whole analysis, CDC data was partitioned into four sets; feature selection data, training data, validation data and test data. Feature selection data was used exclusively for feature variable analysis, training data for model training, validation data for regularization and early stopping while model training, and test data for final model evaluation along with the NYC data set. To sustain the class distribution of the outcome variable, class-stratified random splits of 10%, 70%, 10% and 10% were used, respectively.

```r
# This honestly is not an ideal way to do it. But it looks like the authors
# used caret::createDataPartition to create their stratified samples to split
# train/test/etc. So here goes.

idx_train <- pregs_pp$outcome %>% createDataPartition(p = 0.7, list = FALSE)
pregs_train <- pregs_pp[idx_train, ]

holdout <- pregs_pp[-idx_train, ]
idx_feature <- holdout$outcome %>% createDataPartition(p = 1/3, list = FALSE)
pregs_feature <- holdout[idx_feature, ]

holdout <- holdout[-idx_feature, ]
idx_val <- holdout$outcome %>% createDataPartition(p = 1/2, list = FALSE)
pregs_val <- holdout[idx_val, ]

pregs_test <- holdout[-idx_val, ]

# Sanity checks
# Should be .7, .1, .1, .1 splits
sapply(list(pregs_train, pregs_feature, pregs_val, pregs_test), nrow)
```

```
## [1] 7602 1087 1085 1084
```

```r
# Should be equal
sum(sapply(list(pregs_train, pregs_feature, pregs_val, pregs_test), nrow))
```

```
## [1] 10858
```

```r
nrow(pregs_pp)
```

```
## [1] 10858
```

```r
# Both should be false
any(duplicated(pregs_pp))
```

```
## [1] FALSE
```

```r
any(duplicated(rbind(pregs_train, pregs_feature, pregs_val, pregs_test)))
```

```
## [1] FALSE
```

```r
table(pregs_val$race.f)
```

```
##
## amerindian      asian      black      white
##        215        308        279        283
```

```r
table(pregs_val$education.f)
```

```
##
##       8th   College Doctorate         HS
##       215       279       308       283
```

Feature analysis, feature dropping

> For the task of predictor variable selection, correlation analysis and univariate analysis were used to determine the final set of variables. In correlation analysis, all possible predictor variable pairs were examined for linear dependency to each other with Pearson correlation coefficient. Because highly correlated predictor variables have the same effect on the dependent variable [8], one of the variables with correlation less than -0.5 or more than 0.5 was excluded. This is based on the definition of moderate correlation [21]. This reduces redundancy of the data and produces more robust models.

Let's work in Python to make the heatmap plot, since seaborn has a nice correlation matrix function, and it'll get us accustomed to working the R dataframe/tibbles and Python/Pandas dataframes.

References:

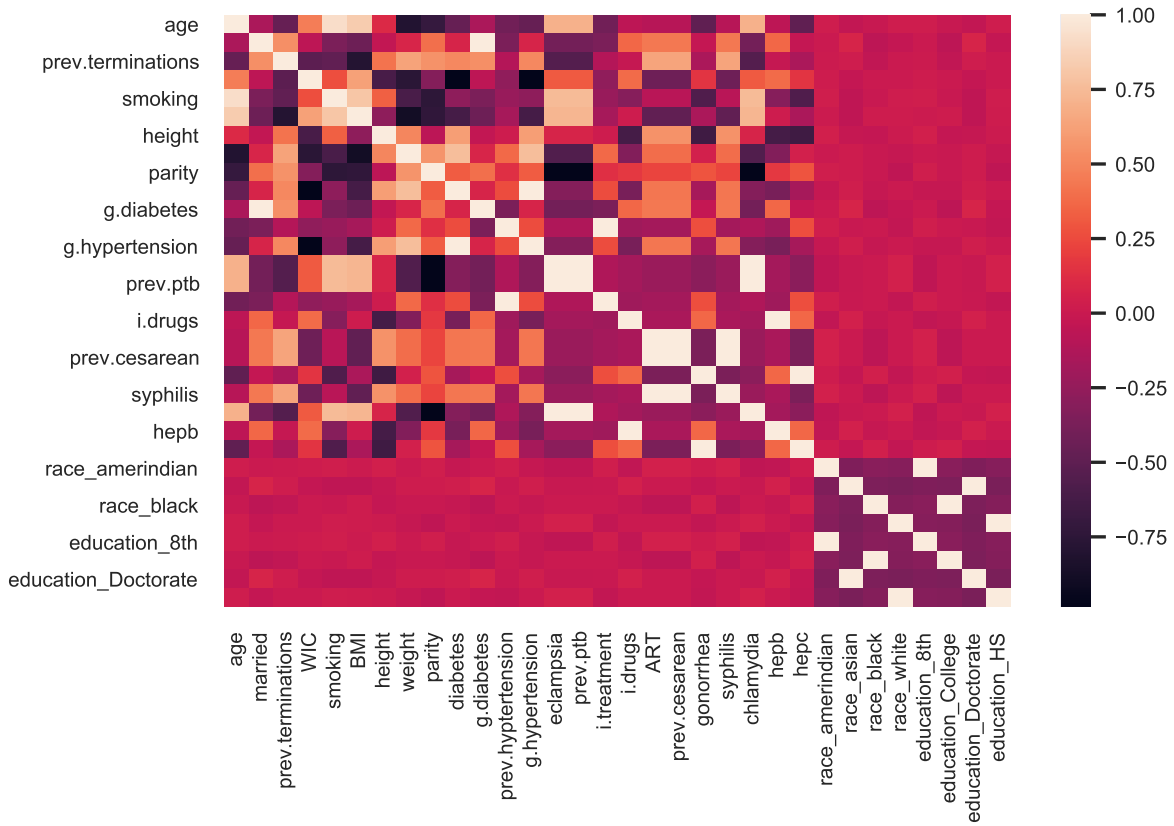
- Start here for working with Python in RStudio: https://rstudio.github.io/reticulate/
- Correlation matrix on a Pandas dataframe: https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html
- Heatmap plot: https://seaborn.pydata.org/generated/seaborn.heatmap.html

```r
import seaborn as sns
import matplotlib.pyplot as plt
from matplotlib import rcParams

# Margins and fonts for our larg tick labels
rcParams.update({'figure.autolayout': True})
plt.gcf().tight_layout()
sns.set(font_scale=0.7)

# Drop outcome columns and the columns we converted to one-hot.
features = r.pregs_feature.drop(columns=['id', 'stillborn', 'g.weeks', 'outcome', 'race.f', 'education.
# Note that the original authors did their correlation plot without one-hots;
# i.e. they treated education and race as factors. It seems extremely tenuous
# to do that with race, and to a lesser degree education, since race definitely
# isn't ordinal (perhaps education is). In any case, here's our analysis with
# respect to the one-hots, which may be more revealing.
```

```
# Pandas built-in function for Rearson correlation
correlations = features.corr()
print(correlations.isna().any())

# Plot
```

```
## age                   False
## married               False
## prev.terminations     False
## WIC                   False
## smoking               False
## BMI                   False
## height                False
## weight                False
## parity                False
## diabetes              False
## g.diabetes            False
## prev.hyptertension    False
## g.hypertension        False
## eclampsia             False
## prev.ptb              False
## i.treatment           False
## i.drugs               False
## ART                   False
## prev.cesarean         False
## gonorrhea             False
## syphilis              False
## chlamydia             False
## hepb                  False
## hepc                  False
## race_amerindian       False
## race_asian            False
## race_black            False
## race_white            False
## education_8th         False
## education_College     False
## education_Doctorate   False
## education_HS          False
## dtype: bool
```

```
sns.heatmap(correlations)
plt.show()
```




```
plt.close()
```

Back to R, we can get the correlation matrix from `py` and filter to see the feature combinations that are highly correlated.

```
corrs <- as.data.frame(py$correlations)
corrs_flat <- lapply(rownames(corrs), function(s) {
  ret <- corrs[s, ] %>%
    select(-c(s)) %>%
    unlist()
  names(ret) <- paste(s, "x", names(ret))
  return(ret)
})
```

```
## Note: Using an external vector in selections is ambiguous.
## i Use `all_of(s)` instead of `s` to silence this message.
## i See <https://tidyselect.r-lib.org/reference/faq-external-vector.html>.
## This message is displayed once per session.
```

```
corrs_flat <- unlist(corrs_flat)
significants <- corrs_flat[corrs_flat > .5 | corrs_flat < -.5]
sort(significants)
```

```
##                parity x eclampsia                  parity x prev.ptb
##                         -0.9840634                         -0.9840634
##                parity x chlamydia                eclampsia x parity
##                         -0.9840634                         -0.9840634
##                 prev.ptb x parity                chlamydia x parity
```

```
##                       -0.9840634                       -0.9840634
##                  WIC x diabetes             WIC x g.hypertension
##                       -0.9797020                       -0.9797020
##                  diabetes x WIC             g.hypertension x WIC
##                       -0.9797020                       -0.9797020
##                    BMI x weight                     weight x BMI
##                       -0.8837906                       -0.8837906
##                    age x weight                     weight x age
##                       -0.8030473                       -0.8030473
##          prev.terminations x BMI          BMI x prev.terminations
##                       -0.7950861                       -0.7950861
##                    WIC x weight                     weight x WIC
##                       -0.7623981                       -0.7623981
##                 smoking x parity                  parity x smoking
##                       -0.7488168                       -0.7488168
##                    BMI x parity                     parity x BMI
##                       -0.7317238                       -0.7317238
##                    age x parity                     parity x age
##                       -0.7014069                       -0.7014069
##                height x gonorrhea                   height x hepc
##                       -0.6511256                       -0.6511256
##                gonorrhea x height                   hepc x height
##                       -0.6511256                       -0.6511256
##                   BMI x diabetes             BMI x g.hypertension
##                       -0.6269853                       -0.6269853
##                   diabetes x BMI             g.hypertension x BMI
##                       -0.6269853                       -0.6269853
##                  height x i.drugs                    height x hepb
##                       -0.6198209                       -0.6198209
##                  i.drugs x height                    hepb x height
##                       -0.6198209                       -0.6198209
##                    WIC x height                     height x WIC
##                       -0.6060979                       -0.6060979
##                  smoking x weight                  weight x smoking
##                       -0.6034102                       -0.6034102
##                smoking x gonorrhea                  smoking x hepc
##                       -0.5657157                       -0.5657157
##                gonorrhea x smoking                  hepc x smoking
##                       -0.5657157                       -0.5657157
##                weight x eclampsia                weight x prev.ptb
##                       -0.5606737                       -0.5606737
##                weight x chlamydia               eclampsia x weight
##                       -0.5606737                       -0.5606737
##                 prev.ptb x weight               chlamydia x weight
##                       -0.5606737                       -0.5606737
##     prev.terminations x eclampsia      prev.terminations x prev.ptb
##                       -0.5495299                       -0.5495299
##     prev.terminations x chlamydia     eclampsia x prev.terminations
##                       -0.5495299                       -0.5495299
##      prev.ptb x prev.terminations     chlamydia x prev.terminations
##                       -0.5495299                       -0.5495299
##           prev.terminations x WIC           WIC x prev.terminations
##                       -0.5058598                       -0.5058598
##      prev.terminations x diabetes prev.terminations x g.hypertension
```

```
##                         0.5114905                         0.5114905
##      diabetes x prev.terminations g.hypertension x prev.terminations
##                         0.5114905                         0.5114905
##    prev.terminations x g.diabetes    g.diabetes x prev.terminations
##                         0.5374090                         0.5374090
##       married x prev.terminations       prev.terminations x married
##                         0.5374090                         0.5374090
##        prev.terminations x parity        parity x prev.terminations
##                         0.5510487                         0.5510487
##                       height x ART            height x prev.cesarean
##                         0.5573510                         0.5573510
##                  height x syphilis                       ART x height
##                         0.5573510                         0.5573510
##            prev.cesarean x height                 syphilis x height
##                         0.5573510                         0.5573510
##                    weight x parity                   parity x weight
##                         0.5626484                         0.5626484
##                  height x diabetes           height x g.hypertension
##                         0.6061782                         0.6061782
##                  diabetes x height           g.hypertension x height
##                         0.6061782                         0.6061782
##                         WIC x BMI                         BMI x WIC
##                         0.6220839                         0.6220839
##        prev.terminations x weight        weight x prev.terminations
##                         0.6340326                         0.6340326
##           prev.terminations x ART prev.terminations x prev.cesarean
##                         0.6372161                         0.6372161
##      prev.terminations x syphilis           ART x prev.terminations
##                         0.6372161                         0.6372161
## prev.cesarean x prev.terminations      syphilis x prev.terminations
##                         0.6372161                         0.6372161
##                    age x eclampsia                    age x prev.ptb
##                         0.7008379                         0.7008379
##                    age x chlamydia                   eclampsia x age
##                         0.7008379                         0.7008379
##                     prev.ptb x age                   chlamydia x age
##                         0.7008379                         0.7008379
##                    BMI x eclampsia                    BMI x prev.ptb
##                         0.7286722                         0.7286722
##                    BMI x chlamydia                   eclampsia x BMI
##                         0.7286722                         0.7286722
##                     prev.ptb x BMI                   chlamydia x BMI
##                         0.7286722                         0.7286722
##                smoking x eclampsia                smoking x prev.ptb
##                         0.7473635                         0.7473635
##                smoking x chlamydia               eclampsia x smoking
##                         0.7473635                         0.7473635
##                 prev.ptb x smoking               chlamydia x smoking
##                         0.7473635                         0.7473635
##                  weight x diabetes           weight x g.hypertension
##                         0.7612271                         0.7612271
##                  diabetes x weight           g.hypertension x weight
##                         0.7612271                         0.7612271
##                      smoking x BMI                     BMI x smoking
```

```
##                          0.7988685                          0.7988685
##                          age x BMI                          BMI x age
##                          0.8346416                          0.8346416
##                      age x smoking                      smoking x age
##                          0.9275105                          0.9275105
##               married x g.diabetes          diabetes x g.hypertension
##                          1.0000000                          1.0000000
##               g.diabetes x married   prev.hyptertension x i.treatment
##                          1.0000000                          1.0000000
##          g.hypertension x diabetes              eclampsia x prev.ptb
##                          1.0000000                          1.0000000
##              eclampsia x chlamydia              prev.ptb x eclampsia
##                          1.0000000                          1.0000000
##               prev.ptb x chlamydia   i.treatment x prev.hyptertension
##                          1.0000000                          1.0000000
##                     i.drugs x hepb               ART x prev.cesarean
##                          1.0000000                          1.0000000
##                     ART x syphilis               prev.cesarean x ART
##                          1.0000000                          1.0000000
##           prev.cesarean x syphilis                  gonorrhea x hepc
##                          1.0000000                          1.0000000
##                     syphilis x ART          syphilis x prev.cesarean
##                          1.0000000                          1.0000000
##              chlamydia x eclampsia              chlamydia x prev.ptb
##                          1.0000000                          1.0000000
##                     hepb x i.drugs                 hepc x gonorrhea
##                          1.0000000                          1.0000000
##    race_amerindian x education_8th   race_asian x education_Doctorate
##                          1.0000000                          1.0000000
##     race_black x education_College          race_white x education_HS
##                          1.0000000                          1.0000000
##    education_8th x race_amerindian     education_College x race_black
##                          1.0000000                          1.0000000
##   education_Doctorate x race_asian          education_HS x race_white
##                          1.0000000                          1.0000000
```

Super wonky right now, but that seems to be just due to our synthetic data.

> The correlation results in Fig. 1 show that mothers BMI (f8) and weight (f10) in pounds were highly correlated (0.94), which makes sense because in the BMI formula BMI=weight(Lb)height(in.)2*703, (3) weight is the numerator. Because of this, weight was chosen to be excluded. Infertility drugs and assisted reproductive technology (ART) use were correlated to infertility treatment (0.68 and 0.67). This is also to be expected, because they are alternative forms of infertility treatment. Figure 2 shows that observations marked for infertility drugs and ART are always a member of the set of infertility treatment. The presence of 10,660 observations (0.08% of all study data) where treatment is marked but drugs or ART use are not suggests either the use of other undocumented infertility treatment procedures such as myomectomy surgeries [18], incomplete documentation in some data collection areas, poor data quality or a combination of the three. Because the data is de-identified, we can only speculate the underlying effect, so the three infertility-related variables were included in the set of features. No other significant correlations were found, i.e. less than -0.5 or more than 0.5.

So let's exclude weight.

```
pregs_train_select <- pregs_train %>% select(-weight)
pregs_val_select <- pregs_val %>% select(-weight)
```

```
pregs_test_select <- pregs_test %>% select(-weight)
```

Model details

Putting this here for reference!

> R (v. 3.5.1) and Python (v. 3.6.9) were used as tools for statistical analysis and modelling. In addition to base packages, R package readr (v. 1.1.1) was used for reading the data set text files [32] … . Several Python packages were used, scipy (v. 1.3.1) [10] and pandas (v. 0.25.0) for data management, and scikit-learn (v. 0.21.2) [24] for logistic regression. … For ML modelling, tensorflow (v. 1.14.0) in conjunction with keras (v. 2.2.4) was used for neural networks [2, 5]. A gradient boosting decision tree was implemented using the lightgbm (v. 2.2.1) package [12].
>
> Logistic regression (LR),gradient boosting decision tree (GBDT) and two artificial neural network (ANN) models were used in this study.

Logistic regression:

> LR will serve as a baseline for the more complex algorithms due to its simplicity and robustness. L2-regularizied logistic regression with limited-memory Broyden–Fletcher–Goldfarb–Shanno (BFGS) parameter optimization was used. Tolerance for stopping criteria was set to 1.0e−4. Regularization strength C was set to 1.0. The optimal maximum number of iterations was found to be 100.

Gradient boosting tree:

> The lightgbm (LGBM) version of GBDT algorithm was chosen for our study. … For modelling, after iterative experimentation the number of leaves was set to 48, minimal number of data observations in one leaf to 500, maximum depth of the tree model was not restricted, shrinkage rate was set to 0.001, feature and bagging fractions were set to 1 and boosting algorithm was chosen to be Gradient Boosting Decision Tree. Maximum iterations was set to 2000, and early stopping after 500 iterations was used and the used metric for performance was AUC. Different outcomes have clinically significant false positive rates based on incidence. True positive rates in those false positive rates could also be used as a metric for performance, however initial experimentation showed that there were no significant changes in using them over AUC.

Neural nets number one:

> For ANN, the first model was a Leaky ReLU-based deep two-layer feed-forward neural network that we have previously shown to perform well in the risk prediction task of Down's syndrome [15].

Neural nets number two:

> The second was a deep feed-forward self-normalizing neural network based on the scaled exponential linear units (SELU) activation function, which has been demonstrated to achieve superior performance to other feed-forward neural networks. … four hidden layers were selected for the SELU network instead of two that was used in our previously published ANN. The number of hidden nodes per layer was set to the number of input variables; all of them contained the SELU activation function. Alpha node dropout amount in these nodes was set to 15%. LeCun normal weight initialization was used. Adam gradient descent optimization with 0.001 learning rate was used for updating weights. Sigmoid activation function was utilized as the final node for binary classification. 10 epochs with a batch size of 256 was tested to be optimal.

Other important notes:

- "For all the case classes of late stillbirth, early stillbirth and PTB, binary classifiers of normal pregnancy vs. case were constructed." I.e. for each method they trained FOUR binary classifiers instead of multiclass classifiers.

- "Because of the class unbalance, class weights w were calculated from the training data set with w=s/(c * f(y)) where s is number of samples, c is the number of different classes and f(y) is the frequency of classes in data labels y."
- "Folded cross validation was determined to not be necessary with the data of this size."
- The authors also built average- and weighted-average ensemble models. "All possible [weighted average] weight combinations were calculated with exhaustive grid search when the objective function was maximizing prediction AUC, with the constraint that the result vector of non-negative values add up to one, i.e. 100%."

Save the feature-selected data

```
# Note - ensure the working directory is top-level (fair-child), not "src"
write.csv(pregs_train_select, "data/sim/pregsim_train.csv", row.names = FALSE)
write.csv(pregs_val_select, "data/sim/pregsim_val.csv", row.names = FALSE)
write.csv(pregs_test_select, "data/sim/pregsim_test.csv", row.names = FALSE)
```