Review Session 3

2024-03-28

Introducing education acceptance

```
require(tidyverse)
```

```
## Loading required package: tidyverse
```

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## Warning: package 'purrr' was built under R version 4.3.3
```

```
## Warning: package 'stringr' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✓ dplyr     1.1.2     ✓ readr     2.1.4
## ✓ forcats   1.0.0     ✓ stringr   1.5.1
## ✓ ggplot2   3.5.0     ✓ tibble    3.2.1
## ✓ lubridate 1.9.2     ✓ tidyr     1.3.0
## ✓ purrr     1.0.2
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts t
o become errors
```

```
ad <- read_rds('https://github.com/jbisbee1/DS1000_S2024/raw/main/data/admit_data.rds')

# Simple Variable Importance
mCost <- glm(yield ~ net_price,ad,family = binomial(link = 'logit'))
summary(mCost)
```

```
##
## Call:
## glm(formula = yield ~ net_price, family = binomial(link = "logit"),
##     data = ad)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  8.497e-01  7.902e-02  10.752   <2e-16 ***
## net_price   -4.078e-06  2.965e-06  -1.375    0.169
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2689.5  on 2149  degrees of freedom
## Residual deviance: 2687.6  on 2148  degrees of freedom
## AIC: 2691.6
##
## Number of Fisher Scoring iterations: 4
```

```
mVisit <- glm(yield ~ visit,ad,family = binomial)
summary(mVisit)
```

```
##
## Call:
## glm(formula = yield ~ visit, family = binomial, data = ad)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.59295    0.05868  10.106  < 2e-16 ***
## visit        0.43299    0.09630   4.496 6.91e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2689.5  on 2149  degrees of freedom
## Residual deviance: 2668.9  on 2148  degrees of freedom
## AIC: 2672.9
##
## Number of Fisher Scoring iterations: 4
```

```
mLegacy <- glm(yield ~ legacy,ad,family = binomial)
summary(mLegacy)
```

```
##
## Call:
## glm(formula = yield ~ legacy, family = binomial, data = ad)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.58187    0.05343  10.890  < 2e-16 ***
## legacy       0.68324    0.11021   6.199 5.67e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2689.5  on 2149  degrees of freedom
## Residual deviance: 2648.6  on 2148  degrees of freedom
## AIC: 2652.6
##
## Number of Fisher Scoring iterations: 4
```

```
# Calculate AUC
require(tidymodels)
```

```
## Loading required package: tidymodels
```

```
## Warning: package 'tidymodels' was built under R version 4.3.3
```

```
## ── Attaching packages ─────────────────────────────────────── tidymodels 1.1.1 ──
## ✓ broom        1.0.5      ✓ rsample      1.2.0
## ✓ dials        1.2.1      ✓ tune         1.1.2
## ✓ infer        1.0.6      ✓ workflows    1.1.4
## ✓ modeldata    1.3.0      ✓ workflowsets 1.0.1
## ✓ parsnip      1.2.0      ✓ yardstick    1.3.0
## ✓ recipes      1.0.10
```

```
## Warning: package 'dials' was built under R version 4.3.3
```

```
## Warning: package 'scales' was built under R version 4.3.3
```

```
## Warning: package 'infer' was built under R version 4.3.3
```

```
## Warning: package 'modeldata' was built under R version 4.3.3
```

```
## Warning: package 'parsnip' was built under R version 4.3.3
```

```
## Warning: package 'recipes' was built under R version 4.3.3
```

```
## Warning: package 'rsample' was built under R version 4.3.3
```

```
## Warning: package 'tune' was built under R version 4.3.3
```

```
## Warning: package 'workflows' was built under R version 4.3.3
```

```
## Warning: package 'workflowsets' was built under R version 4.3.3
```

```
## Warning: package 'yardstick' was built under R version 4.3.3
```

```
## ── Conflicts ─────────────────────────────────────── tidymodels_conflicts() ──
## ✖ scales::discard() masks purrr::discard()
## ✖ dplyr::filter()   masks stats::filter()
## ✖ recipes::fixed()  masks stringr::fixed()
## ✖ dplyr::lag()      masks stats::lag()
## ✖ yardstick::spec() masks readr::spec()
## ✖ recipes::step()   masks stats::step()
## • Learn how to get started at https://www.tidymodels.org/start/
```

```
toEval <- ad %>%
  mutate(prob_yield_cost = predict(mCost,type = 'response'),
         prob_yield_visit = predict(mVisit,type = 'response'),
         prob_yield_legacy = predict(mLegacy,type = 'response'),
         yield = factor(yield,levels = c('1','0')))

roc_auc(toEval,yield,prob_yield_cost)
```

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.533
```

```
roc_auc(toEval,yield,prob_yield_visit)
```

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.551
```

```
roc_auc(toEval,yield,prob_yield_legacy)
```

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.566
```

```
colnames(ad)
```

```
##  [1] "ID"          "income"      "sat"         "gpa"         "visit"
##  [6] "legacy"      "registered"  "sent_scores" "distance"    "tuition"
## [11] "need_aid"    "merit_aid"   "net_price"   "yield"
```

```
mFull <- glm(yield ~ income + sat + gpa + visit + legacy + registered + sent_scores + di
stance + need_aid + merit_aid + net_price,ad,family = binomial)
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(mFull)
```

```
##
## Call:
## glm(formula = yield ~ income + sat + gpa + visit + legacy + registered +
##     sent_scores + distance + need_aid + merit_aid + net_price,
##     family = binomial, data = ad)
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -8.302e+00  1.874e+00  -4.431 9.39e-06 ***
## income       4.876e-05  6.026e-06   8.091 5.92e-16 ***
## sat         -4.962e-04  1.361e-03  -0.365 0.715448
## gpa          1.768e+00  3.372e-01   5.243 1.58e-07 ***
## visit        2.839e-01  1.357e-01   2.093 0.036387 *
## legacy       5.093e-01  1.537e-01   3.313 0.000923 ***
## registered   4.592e-01  1.317e-01   3.486 0.000490 ***
## sent_scores  7.301e-01  1.790e-01   4.078 4.53e-05 ***
## distance    -1.517e-03  3.442e-04  -4.405 1.06e-05 ***
## need_aid    -1.387e-05  2.338e-05  -0.593 0.553138
## merit_aid   -5.252e-06  2.036e-05  -0.258 0.796466
## net_price   -6.240e-05  2.208e-05  -2.826 0.004706 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2689.5  on 2149  degrees of freedom
## Residual deviance: 1454.6  on 2138  degrees of freedom
## AIC: 1478.6
##
## Number of Fisher Scoring iterations: 8
```

```
toEval <- ad %>%
  mutate(prob_yield_full = predict(mFull,type = 'response'),
         yield = factor(yield,levels = c('1','0')))

roc_auc(toEval,yield,prob_yield_full)
```

```
## # A tibble: 1 × 3
##   .metric .estimator .estimate
##   <chr>   <chr>          <dbl>
## 1 roc_auc binary         0.915
```

```
# Cross validation
set.seed(123)
cvRes <- NULL
for(i in 1:100) {
  inds <- sample(1:nrow(ad),size = round(nrow(ad)*.6),replace = F)
  train <- ad %>% slice(inds)
  test <- ad %>% slice(-inds)

  # Train the model
  mFull <- glm(yield ~ income + sat + gpa + visit + legacy + registered + sent_scores +
distance + need_aid + merit_aid + net_price,train,family = binomial)

  # Predicting model on test data
  toEval <- test %>%
  mutate(prob_yield_full = predict(mFull,newdata = test,type = 'response'),
         yield = factor(yield,levels = c('1','0')))

  # Evaluating model performance
  cvRes <- cvRes %>%
    bind_rows(roc_auc(toEval,yield,prob_yield_full) %>%
                mutate(cvInd = i))
}
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
cvRes %>%
  summarise(mean_auc = mean(.estimate))
```

```
## # A tibble: 1 × 1
##   mean_auc
##      <dbl>
## 1    0.912
```

Random Forest for VIMP

```
require(ranger)
```

```
## Loading required package: ranger
```

```
## Warning: package 'ranger' was built under R version 4.3.3
```

```
mRanger <- ranger(yield ~ income + sat + gpa + visit + legacy + registered + sent_scores
+ distance + need_aid + merit_aid + net_price,
                  ad,
                  importance = 'permutation')

data.frame(vimp = mRanger$variable.importance,
           variable = names(mRanger$variable.importance)) %>%
  ggplot(aes(x = vimp,y = reorder(variable,vimp))) +
  geom_bar(stat = 'identity')
```