PLSC 597 - Machine Learning: Assignment 1

Introduction and Replication - Question 1

Harvard Dataverse Link - https://doi.org/10.7910/DVN/OSSD8O
JSTOR Stable Link to Paper - https://www.jstor.org/stable/42940606

Introduction - Is Immigration a Racial Issue? Anglo Attitudes on Immigration Policies in a Border County.

- The purpose of the study is to assess the association between Anglo aversion to Latinos and a variety of other factors including: physical proximity to Latinos, contact with ethnic minorities, and expressed preferences for immigration policies.
- The underlying hypothesis argues immigration policy preferences are strongly influenced by racial resentment rather than other considerations like economic anxiety.
- Other hypotheses argue Anglos living in neighborhoods with larger proportions of Latinos will harbor more restrictive attitudes on immigration, but Anglos who interact more frequently with minorities will harbor less restrictive attitudes on immigration.

Data and Methods

- Data was collected through a telephone survey using random-digit-dial procedures in San Diego County, California in 2005 - 2006. Data were weighted in all regressions to represent San Diego County demographic characteristics for Anglos based on US Census estimates. No significant differences appeared in conclusions when analyses were replicated using unweighted values.
- **Dependent Variable**: The dependent variable of interest was survey respondents' answers to the amnesty question: "As you may know, in 1986 the US Congress passed the Immigration Reform and Control Act, which granted amnesty to nearly 2 million persons who had lived continuously in this country for four or more years without proper documentation. This amnesty law allowed these immigrants to remain here as permanent residents and to apply for US citizenship. At this time, do you think repeating this amnesty program would be a good thing?" 0 = bad idea, 1 = good idea.
- **Independent Variables**
  - Respondent aversion to Latinos (attitude about Latinos): Measured using the Bogardus scale which is a composite index to detect racial attitudes. Recoded as a dummy variable where 1 means aversion was detected and 0 means no aversion was detected.
  - Latino context (concentration in the same Census tract): The natural log of the percent of Latino residents within each respondent's Census tract.
  - Reported contact with minorities: Composite scale summarizing interactions respondents had with Latinos.
- **Controls**
  - Personal financial situation; "In terms of your personal economic situation, would you say that it has improved, remained the same, or gotten worse over the past 12 months?"
  - Family Income.
  - Age.
  - Education.
  - Gender.
  - Political Ideology: "Would you consider yourself conservative, moderate, or liberal?" Liberals were coded as low and conservatives as high.
- **Codebook**

  - Amnesty = amnesty
  - Latino Aversion = dishis2
  - Latino context = pcthis2
  - Latino contact = contact
  - Personal financial situation = retecon
  - Political ideology = idea
  - Education = edu2
  - Famliy Income = income
  - Age = age
  - Sex = male
- **Results**
  - Age, gender, and personal financial situation are not statistically significant. In contrast to the their hypothesis, Latino aversion was not statistically significant either.
  - Increased minority contact was positively associated with amnesty however it was only marginally significant.
  - Decreased income was negatively associated with amnesty however it was only marginally significant.
  - Latino context was negatively and statistically significantly associated with amnesty meaning the more Latinos living in your Census tract the more likely you were to view amnesty as a bad thing.
  - Increased education was positively and statistically significantly associated with amnesty.
  - As one became more conservative, the probability of supporting amnesty decreased. Highly statistically significant.

Packages

```
library(mlr)
library(here)
library(dplyr)
library(foreign)
```

Read in the data

```
ssq <- foreign::read.dta(here("hw1/data/ssq.dta"))
```

Light data cleaning

Select only relevant columns for the regression and remove all rows with a missing value on any of the variables. Turn all variables into numeric values (to match the results from the paper).

```
X2 <-
    ssq %>%
    select(amnesty, dishis2, pcthis2, contact, retecon, idea, edu2, income2,
           age, male, wteth) %>%
    filter(across(everything(), ~!is.na(.x))) %>%
    mutate(across(everything(), as.numeric)) %>%
    mutate(amnesty = as.factor(amnesty))
```

Estimate regression

```
# Make the learning task
amnestyTaskPaper <- makeClassifTask(data = select(X2, -wteth),
                                    target = "amnesty",
                                    weights = X2$wteth)
```

```r
# Create a logistic regression learner
logRegLearner <- makeLearner("classif.logreg", predict.type = "prob")

# Train or fit the model on the whole dataset
amnestyModelPaper <- train(logRegLearner, amnestyTaskPaper)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```r
summary(amnestyModelPaper$learner.model)
```

```
##
## Call:
## stats::glm(formula = f, family = "binomial", data = getTaskData(.task,
##     .subset), weights = .weights, model = FALSE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4768  -0.9430  -0.4715   0.9180   2.6897
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)  1.816576   0.732098   2.481   0.0131 *
## dishis2     -0.549234   0.403517  -1.361   0.1735
## pcthis2     -2.359947   0.973049  -2.425   0.0153 *
## contact      0.083985   0.044441   1.890   0.0588 .
## retecon     -0.233547   0.158242  -1.476   0.1400
## idea        -0.707595   0.163012  -4.341 1.42e-05 ***
## edu2         0.290774   0.131775   2.207   0.0273 *
## income2     -0.128457   0.070815  -1.814   0.0697 .
## age          0.008602   0.006670   1.290   0.1972
## male        -0.345788   0.225834  -1.531   0.1257
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 533.15  on 391  degrees of freedom
## Residual deviance: 484.58  on 382  degrees of freedom
## AIC: 446.5
##
## Number of Fisher Scoring iterations: 4
```

Question 2

Create test and train

```r
set.seed(420)

# Add in new variables
X2new <- X2 %>% mutate(ideaF = as.factor(idea))

# Make test and train sets
train <- X2new %>% sample_frac(0.70)
test <- anti_join(X2new, train)
```

TABLE 3

Ordered and Binary Logistic Regressions of Immigration Policy Preferences on Respondents' Latino Aversion and Selected Predictors Among Anglos, 2005–2006[a]

Predictor	Legal Immigration	Mexican Immigration	Amnesty
Latino aversion	−0.955**	−1.734**	−0.549
	(0.357)	(0.355)	(0.403)
Latino context	−1.671#	−1.332	−2.359*
	(0.890)	(0.855)	(0.973)
Minority contact	0.057	−0.021	0.084#
	(0.043)	(0.039)	(0.044)
Economic evaluation	−0.202	−0.319*	−0.233
	(0.147)	(0.141)	(0.158)
Political ideology	−0.535**	−0.547**	−0.707**
	(0.155)	(0.145)	(0.163)
Education	0.443**	0.249*	0.290*
	(0.127)	(0.117)	(0.131)
Income	−0.026	0.025	−0.128#
	(0.069)	(0.063)	(0.070)
Age	0.003	0.011#	0.008
	(0.006)	(0.006)	(0.006)
Male	0.176	0.019	−0.345
	(0.215)	(0.199)	(0.225)
Cutpoint 1	−1.701	−1.761	1.470
Cutpoint 2	0.819	0.682	
Chi-square	52.138	60.649	48.567
Cox & Snell R^2	0.140	0.140	0.116
N	347	402	392

[a]Numbers in cells are ordered (legal and Mexican immigration) and binary logistic (amnesty) regression coefficients, associated standard errors, and two-tailed probabilities.

#$p<0.10$; * = $p<0.05$; ** = $p<0.01$.

NOTE: Legal and Mexican immigration are derived from policy preferences for increased (3), current levels (2), or decreased immigration (1). Amnesty is derived from preferences for repeating the 1986 amnesty program, a good thing (1) or bad thing (0). List-wise deletion was used for analysis.

Figure 1: Regression Results from Paper

```
## Joining, by = c("amnesty", "dishis2", "pcthis2", "contact", "retecon", "idea", "edu2", "income2", "ag
```

```
paperVars <- c("amnesty", "dishis2", "pcthis2", "contact", "retecon", "idea",
               "edu2", "income2", "age", "male", "wteth")
newVars <- c("amnesty", "pcthis2", "ideaF", "edu2", "wteth")

trainPaper <- train %>% select(all_of(paperVars))
testPaper <- test %>% select(all_of(paperVars))
trainNew <- train %>% select(all_of(newVars))
testNew <- test %>% select(all_of(newVars))
```

Establish cross-validation performance of the model from the paper and the new model

```
# stratified sampling, 10 folds, repeat 200 times
cv <- makeResampleDesc(method = "RepCV", folds = 10, reps = 200, stratify = T)

# Create a task for training a model as specified by the paper
amnestyTaskPaper <- makeClassifTask(data = select(trainPaper, -wteth),
                                    target = "amnesty",
                                    weights = trainPaper$wteth)

# Apply the specified cross validation procedure to the model from the paper
paperTrainCv <- resample("classif.logreg", amnestyTaskPaper, resampling = cv)

# Create a task for training the newly specified model
amnestyTaskNew <- makeClassifTask(data = select(trainNew, -wteth),
                                  target = "amnesty",
                                  weights = trainNew$wteth)

# Apply the specified cross validation procedure to the newly specified model
newTrainCv <- resample("classif.logreg", amnestyTaskNew, resampling = cv)
```

Performance of model from paper (measured through cross validation): 0.3617106
Performance of newly specified model (measured through cross validation): 0.3544603

The combination of treating political ideology as a categorical variable and only including those variables which were found to be statistically significant in the paper (education, political ideology, and Latino context) boosts predictive performance by almost 1% as measured by accuracy.

I suspect I am moving away from the true data generating model. Mostly I am surprised aversion to Latinos wasn't significant and dropping it increases predictive performance (if only slightly). I suspect because the sample size is relatively small and because there aren't many people who reported aversion that these are the reasons the variable didn't reach significance and didn't add much in terms of predictive validity.

Question 3

```
# Train the paper model and my model
paperModelTrain <- train(logRegLearner, amnestyTaskPaper)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
newModelTrain <- train(logRegLearner, amnestyTaskNew)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
# Make predictions on the test set
predictPaperTest <- as_tibble(predict(paperModelTrain, newdata = testPaper))
predictNewTest <- as_tibble(predict(newModelTrain, newdata = testNew))

#Now, let's compute accuracy rate by comparing the truth and response columns
accuracyPaper <-
    predictPaperTest %>%
    mutate(correct = case_when(truth == response ~ 1,
                               truth != response ~ 0))

accuracyNew <-
    predictNewTest %>%
    mutate(correct = case_when(truth == response ~ 1,
                               truth != response ~ 0))

#divide correct cases by total cases
accPaper <- sum(accuracyPaper$correct) / nrow(accuracyPaper)
accNew <- sum(accuracyNew$correct) / nrow(accuracyNew)
```

Accuracy of model specified in the paper on the test data: 0.6355932
Accuracy of the newly specified model: 0.6525424

The newly specified model performs better on the out-of-sample prediction task. This could mean my model is closer to the true data generating process. I am doubtful this is case though for reasons explicated above mostly having to do again with small sample size issues.

```
# Make the learning task
amnestyTaskNew <-
    makeClassifTask(data = select(X2new, amnesty, pcthis2, ideaF, edu2),
                                  target = "amnesty",
                                  weights = X2new$wteth)

# Train or fit the newly specified model on the whole dataset
amnestyModelNew <- train(logRegLearner, amnestyTaskNew)
```

```
## Warning in eval(family$initialize): non-integer #successes in a binomial glm!
```

```
summary(amnestyModelNew$learner.model)
```

```
##
## Call:
## stats::glm(formula = f, family = "binomial", data = getTaskData(.task,
##     .subset), weights = .weights, model = FALSE)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5688  -0.9613  -0.4725   0.9712   2.4915
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   0.3586     0.4507   0.796  0.42618
## pcthis2      -1.9216     0.9289  -2.069  0.03858 *
## ideaF2       -0.7931     0.2551  -3.109  0.00188 **
## ideaF3       -1.4887     0.3075  -4.841 1.29e-06 ***
## edu2          0.1538     0.1177   1.306  0.19158
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 533.15  on 391  degrees of freedom
## Residual deviance: 497.66  on 387  degrees of freedom
## AIC: 448.56
##
## Number of Fisher Scoring iterations: 4
```

Education loses its significance which is mildly interesting. More interesting is the political ideology variable broken out into categories. Even for somebody who professes to be moderate, there is a large decrease in the likelihood to view amnesty as a good idea. It then nearly doubles for those who profess to be conservative. This, to me, suggests further research into the nuances between political ideology and racial views. Even leaving race aside, the framing of the question may be leading. Liberals may be more likely to view those given amnesty as needy and deserving of it while conservatives viewed it as a unworthy handout in a way independent of their views on race.