#sorrynotsorry

...

Ekaterina Titova
Jonathon Guevara

How likely are the death row inmates to be sorry for the crime they committed?
Kaggle data may help us answer this question.

Project Plan

1. Prepare and explore the data:

- Identify the key variables, concatenate columns with names and IDs
- Assign dummy variables to classification variables such as gender, presence of co-defendants, whether the victim was in personal relationship with a perpetrator
- Use such keywords as “regret”, “sorry”, “forgive” to select the statements containing the key word

2. State and/or refine the questions, form the hypothesis

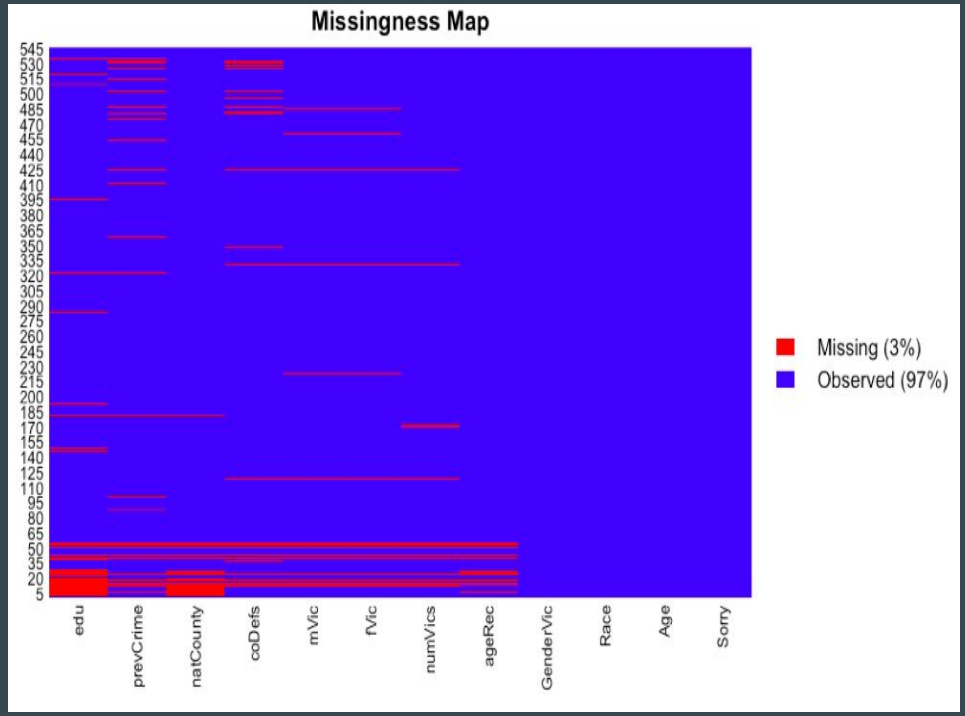
3. Identify a set of approaches to form a prediction Analyze the output of the model:
- Assess the quality of predictive models
- Select the most useful model
- Provide justification for the choices made

4. Methods include but not limited to sqldf library in R, plot.ly to visualize the results

Data Exploration and Preparation

- 3% of missing data with no pattern of missingness
- LOOCV method for resampling
- 545 observations and 9 predictors

  - Response variable “Sorry”
  - Age
  - Race (White, Hispanic, Other)
  - Age when Received
  - Education
  - Native County (binary variable)
  - Previous Crime (binary variable)
  - Co-defendants (binary variable)
  - Gender of the victim (F/M)
  - Number of Victims



Predictor Analysis and Selection:

Age
Race
Education
The history of previous crimes
The presence of co-defendants
Number of Victims

The interactions between predictors:

The history of previous crimes & the presence of co-defendants
The history of previous crimes & the number of victims
Education & Native County
Age & Native County

Models of Interest:

1. GLM
2. GIM Boost
3. SVM
4. PLS
5. MARS
6. KNN
7. LDA
8. GAM
9. LogitBoost

Generalized Linear Model

2 methods have been selected: GLM using train function from caret and stepAIC function with the purpose of exploring interaction terms

AIC for model used for train function = 881
AIC for stepAIC model = 578

Pros:

- Easy to interpret and recognize the significant predictors and interactions
- Provides practical insight in significant interactions

Cons:

- More likely to come with high bias
- The choice of predictors requires careful consideration

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.46857    0.62483   0.750  0.45330
Age          -0.02041    0.01844  -1.107  0.26818
RaceHispanic  0.34884    0.24319   1.434  0.15145
RaceOther   -13.58168  507.46016  -0.027  0.97865
RaceWhite     0.77516    0.20042   3.868  0.00011 ***
ageRec       -0.02117    0.02003  -1.057  0.29040
edu          -0.03939    0.04215  -0.935  0.34999
natCounty1    0.38709    0.17627   2.196  0.02809 *
prevCrime1   -0.05927    0.17355  -0.342  0.73269
coDefs1      -0.29319    0.17566  -1.669  0.09509 .
GenderVicM    0.13005    0.17000   0.765  0.44427
numVics       0.08160    0.10922   0.747  0.45495
---
```

```
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.331308   0.954943  -0.347  0.72864
Age               -0.007657   0.015963  -0.480  0.63147
RaceHispanic       0.335172   0.310765   1.079  0.28079
RaceOther        -13.863656 624.135327  -0.022  0.98228
RaceWhite          0.801947   0.247073   3.246  0.00117 **
edu               -0.149431   0.067820  -2.203  0.02757 *
natCounty1         0.923237   1.477663   0.625  0.53211
prevCrime1         1.543178   0.532648   2.897  0.00377 **
coDefs1            1.544881   0.559413   2.762  0.00575 **
numVics            0.478361   0.212730   2.249  0.02453 *
prevCrime1:coDefs1 -1.324007  0.442679  -2.991  0.00278 **
prevCrime1:numVics -0.984796  0.329754  -2.986  0.00282 **
coDefs1:numVics    -0.588642  0.284835  -2.067  0.03877 *
edu:natCounty1     0.174709   0.103206   1.693  0.09049 .
Age:natCounty1    -0.053027   0.027283  -1.944  0.05194 .
natCounty1:coDefs1 -0.739827  0.456048  -1.622  0.10475
```

Boosted Generalized Linear Model (GLM Boost)

"If linear regression was a Toyota Camry, then gradient boosting would be a UH-60 Blackhawk Helicopter."
Ben Gorman, a Kaggle Master

Pros:

- Provides an frequency based hierarchical view of significant predictors
- Shrinkage function that allows to leveraging the penalty parameter to leverage the predictive accuracy

Cons:

- Concept of negative likelihood may be difficult to interpret

```
Selection frequencies:
   RaceWhite    RaceOther RaceHispanic      coDefs1   natCounty1       ageRec          Age
  0.18666667   0.12000000   0.10666667   0.10666667   0.10000000   0.09333333   0.08666667
         edu   GenderVicM      numVics  (Intercept)   prevCrime1
  0.06000000   0.05333333   0.04666667   0.02000000   0.02000000
```

PLS

- 2 components would be needed for the optimal fit
- The only available model from the dimension reducing family of models

Pros:

- Provides insight into variable importance
- Allows to assess and make a selection of the relevant predictors
- No requirement for a linear relationship between response and predictors

Cons:

- May require a lot of modification for predictors' space to improve the fit of the model

	Overall
RaceWhite	100.000
natCounty1	70.375
coDefs1	38.180
edu	36.004
numVics	18.388
prevCrime1	11.284
ageRec	10.266
Age	8.457
RaceHispanic	6.677
RaceOther	1.214
GenderVicM	0.000

SVM with radial basis function

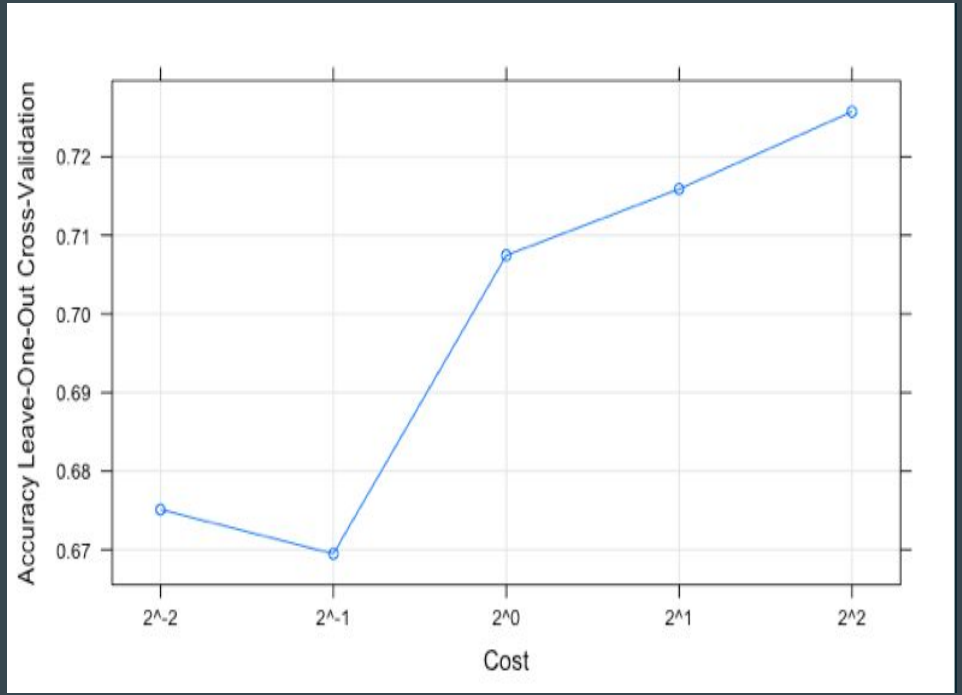
- Used tuneLength = 5 to reduce the computational time
- The optimal accuracy of 72.5% is achieved at the cost = 4

Pros:

- Provides the highest level of accuracy compared to other models
- Works the best with the smaller datasets
- Kernel tricks offer flexibility for designing various models

Cons:

- May be difficult to interpret



Multivariate Adaptive Regression Splines (MARS)

Consistent with GLM model based on selecting Race and Native County as significant predictors

Pros:

- Not as affected by the high collinearity of predictors
- Little to no data preprocessing is required
- Allows clear interpretation of each predictors' impact
- Provides the importance function which ranks the importance of each predictor

Cons:

- Random selection of predictors, which may lead to redundancy

	Overall
RaceWhite	100.00
ageRec	87.53
natCounty1	31.81
Age	21.72
edu	20.08
numVics	0.00
prevCrime1	0.00
GenderVicM	0.00
coDefs1	0.00
RaceHispanic	0.00
RaceOther	0.00

KNN

Classification by having the objects assigned to a class most common of its k nearest neighbors

Pros:

- Easily implemented/interpreted

Cons:

- Computational inefficiency
- High data storage

Accuracy: 66.2% at k=13

	k	Accuracy	Kappa
1	5	0.6582278	0.159562204
2	7	0.6399437	0.082367888
3	9	0.6497890	0.080923235
4	11	0.6540084	0.083705287
5	13	0.6624473	0.096626644
6	15	0.6526020	0.059115676
7	17	0.6399437	0.020914020
8	19	0.6399437	0.007622101
9	21	0.6329114	-0.035662263
10	23	0.6441632	-0.020942943

LDA

Attempts to find linear combinations of the variables to model the differences of the classes of the data

Pros:

- Performs better than logistic regression when response classes are well-defined and when the sample size is small

Cons:

- Strict model assumptions (normality and equal covariances)

Accuracy: 68.2%

```
  parameter  Accuracy      Kappa
1      none 0.6821378 0.08672077
```

Generalized Additive Model

Pros:

- Fits for precision of the model

Cons:

- Smoothing can cause overfitting
- Sacrifices predictability for precision

Accuracy:

- TRUE: 68.1%
- FALSE: 68.2%

	select	method	Accuracy	Kappa
1	FALSE	GCV.Cp	0.6821378	0.1402568
2	TRUE	GCV.Cp	0.6807314	0.1353006

	Overall
RaceWhite	100.000
ageRec	81.370
natCounty1	53.476
coDefs1	30.845
RaceHispanic	17.796
Age	8.713
numVics	3.862
edu	3.826
GenderVicM	2.669
prevCrime1	0.000

Logit Boost

Summary:

Additive logistic model that is implemented as a generalized additive model which employs the the cost functional of logistic regression. The boosting of the model finds sets of weak learners to be used a strong learner in the estimation.

Accuracy: 66.4% at iteration 31

Accuracy <dbl>
0.6468085
0.6170213
0.6638298
0.6553191
0.5808511
0.6255319
0.6021277
0.6000000
0.6085106
0.6319149

Questions of interest

- Based on the nature of the relationship between the victim and the inmate(immediate family, romantic partner vs a cop or immediate family, romantic partner vs a stranger), how likely is the inmate to express regret in his final statement?
- Does a presence of a co-defendant have any impact on expressing regret in the final statement?
- Does a confession always result in a final statement with the words of regret?
- What are the factors influencing those, who chose not to give a final statement?
- Does the number of victims have any relationship with the regret?

Predictor Analysis and Selection:

Age
Race
Education
The history of previous crimes
The presence of co-defendants
Number of Victims

The interactions between predictors:

The history of previous crimes & the presence of co-defendants
The history of previous crimes & the number of victims
Education & Native County
Age & Native County

Variable Importance Comparison

Predictors & their importance	Age	Race	edu	Native County	Age When Received	Previous Crime	Co-defendants	Victim Gender	Number of Victims
Glm	x	X	x	X		x	X		x
Glm boost	x	X	x	X	x	x	X	x	x
PLS	x	X	x	X	x	x	X		x
MARS	X	X	X	x	x				
GAM	X	X	x	X		x	x	x	x

Model Performance Comparison

Models	Accuracy	Kappa	Parameters
GLM	0.683	0.092	3 significant predictors
GLM Boost	0.673	0.037	Mstop = 150
SVM	0.725	0.29	cost=4 and sigma =0.06
PLS	0.682	0.031	ncomp=2
MARS	0.694	0.173	nprune=12
KNN	0.662	0.096	k=13
LDA	0.682	0.086	n/a
GAM	0.681 0.682	0.135 0.140	True False
Logit Boost	0.664		Iteration: 31

and the winners are:

Models	Accuracy	Kappa	Parameters
GLM	0.683	0.092	3 significant predictors
GLM Boost	0.673	0.037	Mstop = 150
SVM	**0.725**	**0.29**	**cost=4 and sigma =0.06**
PLS	0.682	0.031	ncomp=2
MARS	**0.694**	**0.173**	**nprune=12**
KNN	0.662	0.096	k=13
LDA	0.682	0.086	n/a
GAM	0.681 0.682	0.135 0.140	True False
Logit Boost	0.664		Iteration: 31