Project Presentation

AN ANALYSIS THE U.S. GUN CONTROL LOGISTIC CLASSIFICATION PROBLEM

Xin Xu 510060497
Xingyu Zhao 311008720
Guanhong Yu 510318631
Yutong Cai 510135555
Sixiao Chen 510211077
Yuxin Hou 500673599
Yinchu Zhao 510246105

Background

1. Mass gun shooting incidents happened in the U.S
2. Increasingly voice of the people to strengthen gun control laws
3. The debate about gun control policy has always been a controversial topic in the United States.

Dataset Description

1. Our dataset is based on General Social Survey (GSS) which collected societal change and studied the growing complexity of American society
2. We selected 119 features related to gun control independent variable
3. We deleted the columns which contains over 50% missing values and finally got 41 features

GUNLAW	OWNGUN	CHILDS	SIBS	AGE	SEX	RACE	DEGREE	RELIG
<fctr>	<fctr>	<dbl>	<dbl>	<dbl>	<fctr>	<fctr>	<fctr>	<fctr>

Feature engineering

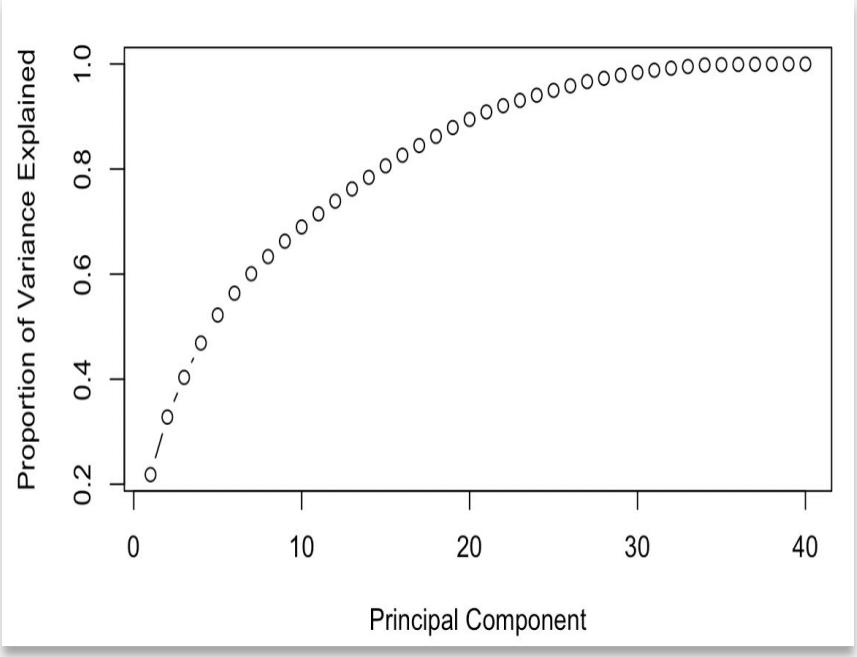
- Delete the missing value in the classification label (Gun law)
- The dataset combines categorical and numeric data
- Deal with these missing values respectively

Data Cleaning

- Clean missing value on respondent's highest degree, total family income and respondent's income features
- About 1% missing value of each feature
- Replace option with eight or more into numeric 8 in some variables like age and earning level

Categorical missing value fill in and dimensional reduction

- Replace with ‘undecided’ and ‘other’ in some categorical features such as, job satisfaction or religion
- Using KNN for filling in numeric missing value
- Apply PCA to do dimensional reduction analysis



Logistic Regression

Introduction
1)A supervised learning classification algorithm

2)binary classification or multivariate classification

3)predict the probability of a target variable

Advantages:
1) Fast speed, suitable for binary classification problems
2) Simple and easy to understand, directly see the weight of each feature
3) The model can be easily updated to absorb new data

Disadvantages:
Its adaptability to data and scenarios is not as strong as that of decision tree algorithm

Process

```r
data <- gun_numeric
set.seed(500673599)
index <- sort(sample(nrow(data), nrow(data)*.7)) # Training set Test set 7:3
training <- data[index,]
testing <-  data[-index,]

glm.fit=glm((as.factor(GUNLAW))~.,data=training,family=binomial(link="logit"))
```

1)After data preprocessing, the data set reserved for GUNLAW column was randomly split into 0.7 training set and 0.3 test set

2）Using the LR algorithm for model training.(The independent variable is GUNLAW column, and the dependent variable is other columns)

3）Use the model to predict the test data

4）Compute Cox-snell goodness of fit and Nagelkerke goodness of fit

5）Evaluation the model

```r
p=predict(glm.fit,testing)#Use the mo
p=exp(p)/(1+exp(p))#Compute the depen
testing$GUNLAW_predicted=1*(p>0.5)#Ad
```

Analysis

1) The results of Cox-Snell goodness of fit and Nagelkerke goodness indicate that the fitting degree of the model is not good enough

2) precision is 0.47,recall is 0.03,F-measure is 0.06,
It shows that the overall accuracy of prediction results is not very good.

3) further feature engineering processs the data.

```
Cox-Snell R2= 0.09968602
```

```
Nagelkerke R2= 0.1503608
[1] 0.4707521
[1] 0.03335965
[1] 0.06230415
```

```
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.677e+01  3.589e+00  -4.673 2.96e-06 ***
CHILDS         4.376e-02  1.424e-02   3.072 0.002125 **
YEAR           8.518e-03  1.822e-03   4.674 2.95e-06 ***
EARNRS         5.018e-02  2.791e-02   1.798 0.072203 .
TEENS          6.243e-02  3.985e-02   1.567 0.117168
OLD1          -6.783e-03  1.682e-03  -4.032 5.52e-05 ***
HOMPOP        -5.350e-02  1.895e-02  -2.823 0.004763 **
SIZE          -7.826e-05  2.416e-05  -3.240 0.001197 **
SEI10         -1.210e-02  3.618e-03  -3.345 0.000824 ***
SEI10INC       8.621e-03  2.512e-03   3.431 0.000600 ***
PASEI10       -1.417e-02  3.599e-03  -3.938 8.22e-05 ***
PASEI10INC     8.308e-03  2.538e-03   3.273 0.001065 **
PRESTG105PLUS  2.310e-03  1.560e-03   1.481 0.138560
PAPRES10      -2.417e-02  8.544e-03  -2.829 0.004668 **
PAPRES105PLUS  1.745e-02  4.497e-03   3.881 0.000104 ***
OWNGUN         3.375e-01  1.465e-02  23.039  < 2e-16 ***
SEX           -7.114e-01  4.636e-02 -15.344  < 2e-16 ***
RACE          -3.187e-01  4.498e-02  -7.087 1.37e-12 ***
DEGREE        -9.202e-02  2.387e-02  -3.855 0.000116 ***
RELIG          2.524e-02  4.481e-03   5.632 1.78e-08 ***
REGION         7.810e-02  8.364e-03   9.338  < 2e-16 ***
INCOME        -4.188e-02  1.026e-02  -4.082 4.46e-05 ***
```

KNN

What is KNN?

The idea is that a sample belongs to a category if most of the k most similar samples in the feature space belong to that category.

Advantages:

1. Simple, easy to understand, easy to implement, no parameter estimation, no training;

2. Suitable for classifying rare events;

Disadvantages:

1. Large amount of calculation, because the distance between each text to be classified and all known samples must be calculated, to obtain its K nearest neighbour points;

KNN Process

1. Prepare and pre-process data

2.Randomly split data into 0.7 training set and 0.3 test set and deal with Null value.

3.train the GUNLAW data by using KNN.

4.Use this models to predict test data.

5.Analysis the results

```
library(caret)

## 75% of the sample size
sample_data= floor(0.7 * nrow(gun))

train_ind <- sample(seq_len(nrow(gun)), size = sample_data)

train <- gun[train_ind, ]
test <- gun[-train_ind, ]


#deal with NA value using knn

missing_data_train <- preProcess(train, method='knnImpute')
missing_data_test <- preProcess(test, method='knnImpute')
train <- predict(missing_data_train, newdata = train)
test <- predict(missing_data_test, newdata = test)

#check if na in train
anyNA(train)
anyNA(test)
```

```
# knn
KNN.fit = train(GUNLAW ~ ., data=train, method="knn",trControl =
trainControl(method = "repeatedcv",repeats = 5))
KNN.fit
```

Analysis

Using confusion Matrix from Caret, we saw that the KNN algorithm was 74.09 percent accurate in this case.

95% Confidences interval is between (0.7303,0.7513)

```
KNN_pre=predict(KNN.fit,test)
KNN_matrix = caret::confusionMatrix(KNN_pre,test$GUNLAW)
KNN_matrix
```

```
Confusion Matrix and Statistics

          Reference
Prediction FAVOR OPPOSE
    FAVOR   4852   1504
    OPPOSE   249    160

               Accuracy : 0.7409
                 95% CI : (0.7303, 0.7513)
    No Information Rate : 0.754
    P-Value [Acc > NIR] : 0.994

                  Kappa : 0.0635

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.95119
            Specificity : 0.09615
         Pos Pred Value : 0.76337
```

Random Forest

- 1.Before adopting the Random Forest classification method, the model was split in to a 0.7 training set and a 0.3 test set in order to perform the algorithm

Process Of The Classification

	FAVOR	OPPOSE	MeanDecreaseAccuracy	MeanDecreaseGini
OWNGUN	19.0974578	57.067124	51.048796	246.49367
CHILDS	14.4539158	-5.045065	11.091616	89.91266
SIBS	7.7280891	-2.795039	5.154816	123.26461
AGE	20.9672710	-10.652834	16.043000	159.75653
SEX	16.3494801	29.082879	30.864352	107.27564
RACE	-0.5326896	15.743154	9.461451	36.49991
DEGREE	15.3661421	-10.248429	14.697023	65.71169
RELIG	7.8869240	5.739904	9.808064	96.43396
PARTYID	6.9347310	11.690022	12.235473	220.33126
REGION	16.2277560	13.530166	20.464970	261.14077
YEAR	15.9352471	7.003612	17.647451	178.93082
EDUC	20.1977702	-14.708297	18.660987	108.38138
PAEDUC	21.1223043	-9.306758	17.478029	141.58000
MAEDUC	15.5556705	-6.569135	10.785460	121.57374
EARNRS	8.3163796	-3.849001	5.892276	56.12034
TEENS	5.6098724	-2.141293	3.968601	36.91014
ADULTS	5.9988338	-1.337775	5.412890	48.50026
INCOME	20.6521874	-8.768962	18.545192	137.95211
RINCOME	33.5709440	-18.468862	27.287665	217.74306
SATJOB	1.1617418	2.565445	2.316267	91.49569
OLD1	19.3590120	-6.556480	16.311010	169.20726
OLD2	19.1537919	-7.751762	14.546991	179.10636
SPEDUC	29.0198136	-10.388810	24.854661	164.71789
HOMPOP	10.1460610	-4.921276	6.531925	79.07688
SIZE	6.9035359	19.219243	16.482582	230.38476
TVHOURS	6.6737813	-4.682660	3.757264	160.54501
REALINC	29.1227399	-15.658793	25.709495	169.83475
REALRINC	28.6989308	-14.050910	26.438791	173.63384
CONINC	30.1534914	-16.653019	27.064824	171.91210
CONRINC	28.8549774	-14.390403	27.130063	172.79119
SEI10	24.6718380	-15.200742	23.719954	156.03730
SEI10EDUC	25.8227477	-14.553846	24.132765	172.16203
SEI10INC	24.7266600	-9.199544	22.961668	172.07507
PASEI10	23.6895133	-12.658383	22.196738	155.49966
PASEI10EDUC	20.7216757	-10.229456	18.201836	160.46460
PASEI10INC	25.0125024	-11.613003	21.500226	163.16078
PRESTG10	22.8477839	-10.472486	21.478408	137.09373
PRESTG105PLUS	24.5295328	-11.061203	23.029149	147.34020
PAPRES10	26.1169911	-12.831532	22.876461	141.25723
PAPRES105PLUS	23.1610773	-10.536573	19.805102	145.77888

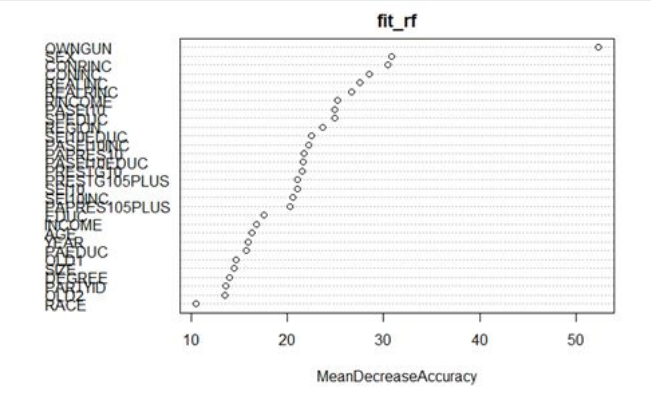
	pred_rf	
	FAVOR	OPPOSE
FAVOR	5037	140
OPPOSE	1400	188

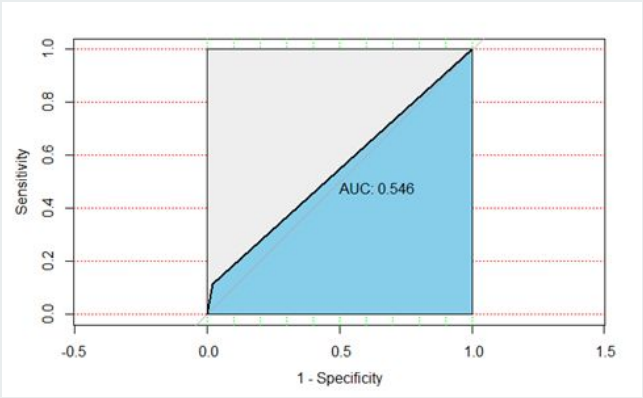
- 500 trees were established which 6 variables tried at each split of the boosting.
- It can be discovered that family income(REALINC,CONINC) are the biggest element that affect people who favors gun law and having an own gun affects the oppose.

```
                Type of random forest: classification
                      Number of trees: 500
No. of variables tried at each split: 6

        OOB estimate of  error rate: 23.27%
Confusion matrix:
       FAVOR OPPOSE class.error
FAVOR  11703    317  0.02637271
OPPOSE  3356    408  0.89160468
```

Classification Analysis



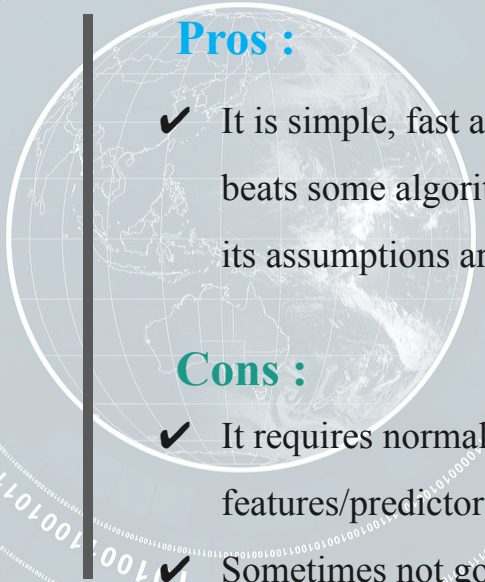


- **The model result in an accuracy at 0.7709 which for all classes of positive and negative predictions, 77.09% were predicted correctly.**
- **AUC around 0.55 which means the model has little issue with the measure of separability**
- **Out of Bag error rate exists at 23.27% as the model used 70% of the data set as training which may cause unpredicted estimation during bootstrap**

LINEAR DISCRIMINANT ANALYSIS(LDA)

LDA

- Malignant or benign
- Making profit or not
- Buy a product or not
- Satisfied customer or not



Pros :

- ✔ It is simple, fast and portable algorithm. It still beats some algorithms (logistic regression) when its assumptions are met.

Cons :

- ✔ It requires normal distribution assumption on features/predictors.
- ✔ Sometimes not good for few categories variables.

LDA

1. prepare data for further analysis by splitting the dataset into 3/7 ratio.
2. nominate the rows selected for train and test data by name the target variable inTrain.
3. Apply lDA model, input target attribute and data set as arguments.
4. Compare the predicted value with test dataset, use confusion matrix to present the goodness of the prediction produced by the lDA model.

```
library(caret)
set.seed(125)
gun_numeric <- data.frame(gun_numeric)
inTrain <- createDataPartition(gun_numeric$GUNLAW, p = 0.7)[[1]]
guntrain <- gun_numeric[ inTrain, ]
guntest  <- gun_numeric[-inTrain, ]
#head(inTrain)
#nrow(guntrain)
#nrow(guntest)
lda.model <- train(GUNLAW ~ . , data = guntrain, method = "lda",
                   trControl = trainControl(method = "repeatedcv", repeats = 5))
lda.model

lda.pred <- predict(lda.model, newdata = guntest)
lda.confusion <- caret::confusionMatrix(lda.pred, guntest$GUNLAW)
lda.confusion
```

LDA

```r
lda.pred <- predict(lda.model, newdata = guntest)
lda.confusion <- caret::confusionMatrix(lda.pred, guntest$GUNLAW)
lda.confusion
```

```
Confusion Matrix and Statistics

          Reference
Prediction FAVOR OPPOSE
    FAVOR   4943   1395
    OPPOSE   205    221

               Accuracy : 0.7635
                 95% CI : (0.7531, 0.7735)
    No Information Rate : 0.7611
    P-Value [Acc > NIR] : 0.33

                  Kappa : 0.1297

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.9602
            Specificity : 0.1368
         Pos Pred Value : 0.7799
         Neg Pred Value : 0.5188
             Prevalence : 0.7611
         Detection Rate : 0.7308
   Detection Prevalence : 0.9370
      Balanced Accuracy : 0.5485

       'Positive' Class : FAVOR
```

1. Comparison results display

01 Logistic Regression

precision is 0.47
recall is 0.03
F-measure is 0.06

02 KNN

Accuracy rate 74.55%
95% confidence interval
is[0.7339,0.7558]

03 Random forest

Accuracy rate 77.09%
AUC is 0.55

LDA

Accuracy rate 76.35%
95% confidence interval
[0.7531,0.7735]

2. Experiment results analysis

Random forest -***Ranked top1 model 77.09%***

Advantage:

Random forest is very stable.

By averaging the decision tree, the risk of overfitting is reduced.

Disadvantage:

It is more complicated than the decision tree algorithm and has a higher computational cost.

Due to their complexity, they require more time to train than other similar algorithms.

LDA-***Ranked top2 model 76.35%***

Advantage:

Prior knowledge experience of the label can be used

Disadvantage:

LDA is not suitable for dimensionality reduction of non-Gaussian samples.

2. Experiment results analysis

Logistic Regression***- Reasons for low accuracy 47%***

Advantage:

The output value between 0 and 1, which has probabilistic meaning.(sigmoid)

Disadvantage:

LR is a linear classifier, so it cannot handle the correlation between features.
When the feature space is large, the performance is not good.
It is prone to under-fitting, and the accuracy is not high.

KNN-***Ranked third model 74.55%***

Advantage:

High accuracy, Not sensitive to outliers
It can be used for classification and regression

Disadvantage:

The amount of calculation is large, especially when the number of features is very large
The interpretability of the KNN model is not strong
Model training takes too long

- Summary

1. The optimal model in this experiment is random forest, which reached 77.00%
2. Understand the characteristics of different classifiers
3. Understand the applicable scenarios of different classifiers
4. Understand the importance of data processing to data analysis