**Project**

**US Congressional Voting Records**

Author

**C Vinod Kumar**

Batch 31

Under the guidance of

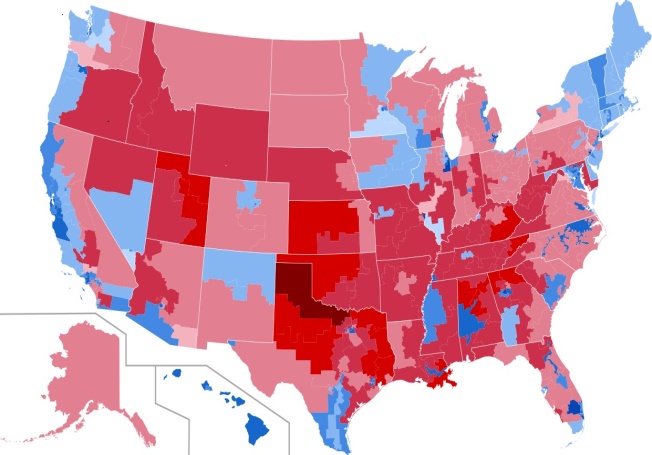
Anjali Gummaraju

INSOFE Research Internship Program

**Introduction:**

Congressional Quarterly Almanac, 98th Congress, 2nd session 1984, Volume XL: Congressional Quarterly Inc. Washington, D.C., 1985.

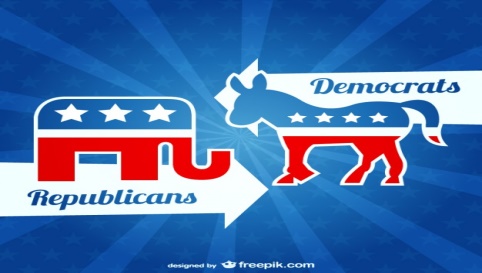
Given data is the official source of information of US Congressional voting on each bill produced in House of representatives. In US has 50 states and 435 congressional districts each district elects a representative to the House of Representatives for a 2-year term. Each year the U.S. Senate and House of Representatives take thousands of votes to approve or reject the bill. In that 16 bills were given for our business case study problem



**Abstract:**

HouseVotes84 dataset data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to YES), voted against, paired against, and announced against (these three simplified to No), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition).

***Objective: “Goal is to learning the Voting Pattern of given data & predict/classify each representative’s party in the house”***



**Congressional Record**

The Congressional Record is the official source of information on recorded floor votes. Votes are printed in the daily Record as they occur on the floor. The votes provide an alphabetical listing of members under “yea,” “nay,” and “not voting” categories and show the overall tally for each category.  However, votes are not identified by party. The Congressional Record Index provides subject access to the votes (under “Votes in Senate” and “Votes in House”).

**Table of contents**:

1. Getting Data
2. Exploring Data
3. Data Visualization
4. Missing Value Imputation
5. Feature Engineering
6. Machine Learning
7. Feature Importance
8. Naïve bayes
9. Logistic Regression
10. Decision Tree
11. Random Forest

**1. Getting Data:**

Given HouseVotes84 dataset in txt type I read data into R using “read. table ( )" command.

**R Code:**

#Reading given txt file into table format

actual\_data <- read.table("house-votes-84.data.txt",sep=",")

Given data frame consists of 435 observations on 17 variables represent - 435 congress members and 16 attributes represents bills and 1 target variable

**2. Exploring data**

It is an classification problem all x and y variables’ are categorical, levels 2 in each variable

Data set characteristics : Multivariate

Data set characteristics : Categorical

Associated task : Classification

Number of instances : 435 instances

Number of attributes : 16 attributes

Number of missing values : 392 missing values

Given data doesn’t a havening variable name to describe, by default V1, V2, V3, V4….V17 assigned as column names

> #checking the summary of data

> summary(actual\_data)

V1 V2 V3 V4 V5 V6 V7 V8

democrat :267 ?: 12 ?: 48 ?: 11 ?: 11 ?: 15 ?: 11 ?: 14

republican:168 n:236 n:192 n:171 n:247 n:208 n:152 n:182

y:187 y:195 y:253 y:177 y:212 y:272 y:239

V9 V10 V11 V12 V13 V14 V15

?: 15 ?: 22 ?: 7 ?: 21 ?: 31 ?: 25 ?: 17

n:178 n:206 n:212 n:264 n:233 n:201 n:170

y:242 y:207 y:216 y:150 y:171 y:209 y:248

V16 V17

?: 28 ?:104

n: 233 n: 62

y: 174 y:269

So, rename column names with below attribute information

**Attribute Information:**

Variable name levels and values

1. Class Name : 2 (democrat, republican)   
2. handicapped-infants : 2 (y,n)   
3. water-project-cost-sharing : 2 (y,n)   
4. adoption-of-the-budget-resolution : 2 (y,n)   
5. physician-fee-freeze : 2 (y,n)   
6. el-salvador-aid : 2 (y,n)   
7. religious-groups-in-schools : 2 (y,n)   
8. anti-satellite-test-ban : 2 (y,n)   
9. aid-to-nicaraguan-contras : 2 (y,n)   
10. mx-missile : 2 (y,n)   
11. Immigration : 2 (y,n)   
12. synfuels-corporation-cutback : 2 (y,n)   
13. education-spending : 2 (y,n)   
14. superfund-right-to-sue : 2 (y,n)   
15. Crime : 2 (y,n)   
16. duty-free-exports : 2 (y,n)   
17. export-administration-act-South-Africa: 2 (y,n)

Each bill represents voting pattern of each member as “Y” (Voted Yes),“N” (Voted No), “?” (Not voted at all)

Replaced “?” with “NA” to understand the how many missing values ,

Data set having “392” missing values i.e., “5.63%” data is having missing values

#Total percentage of missing values in dataset

> (sum(is.na(actual\_data))/(nrow(actual\_data)\*ncol(actual\_data)))\*100 #I found 5.300879% of data missing

[1] 5.300879

> #Total percentage of missing values in each variable of given dataset

> (colSums(is.na(actual\_data))/nrow(actual\_data))\*100

V1 V2 V3 V4 V5 V6 V7

0.000000 2.758621 11.034483 2.528736 2.528736 3.448276 2.528736

V8 V9 V10 V11 V12 V13

3.218391 3.448276 5.057471 1.609195 4.827586 7.126437

V14 V15 V16 V17

5.747126 3.908046 6.436782 23.908046

While replacing “?” as NA only values changes but levels didn’t changed so it is showing as 3 levels so we dropped levels apart from actual levels

#here showing 3 levels inculing "na" so we dropped "na"

> droplevels(actual\_data)->actual\_data

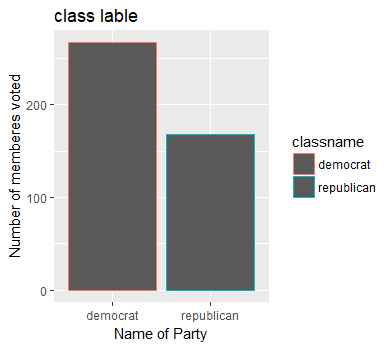
**3. Data visualization:**

Plot to know which party is majority and which party is minority

democrat republican

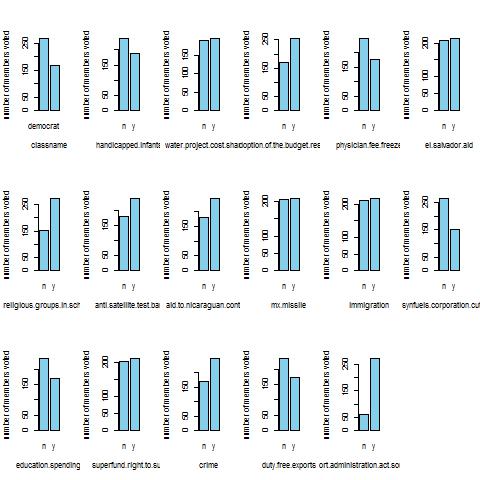
267 168

plot(x = classname,color=classname,xlab="Name of Party",ylab ='Number of memberes voted', data = actual\_data, geom = "auto",main ="class lable")



Here democrats are more compared to republicans

Congressional voting pattern



**4. Missing values handling**

* Given data set contains missing values
* KNN doesn't work to my data set because here hamming distance doesn't work for categorical, So I used Central Imputation to my data set by using method "mode" ,
* Central Imputation by default uses method "median" to replace the missing values

To do imputation divided data with respective to each class name

1. All republican's one data frame
2. All democrats in one data frame

Because Central Imputation replaces missing value with “mode” (whose party votes more) so while imputing it will impute other party votes to missing values

We imputed missing values on each data frame separately and combined using rbind

#spliting data into republic and democratic

republic <- congress\_data[which(congress\_data$classname == 'republican'),]

democrat <- congress\_data[which(congress\_data$classname == 'democrat'),]

> imputed\_congress\_data <- rbind(republic,democrat)

**5. Data Preprocessing**

Dividing training and test data into 80, 20 ratio

#divide into data in to training,test 80,20 proportions

set.seed(5000)

proportion\_data <- sample(seq(1,2),size = nrow(imputed\_congress\_data),replace = TRUE, prob = c(.8, .2))

train\_congress\_data <- imputed\_congress\_data[proportion\_data == 1,]

test\_congress\_data <- imputed\_congress\_data[proportion\_data == 2,]

**6. Model building:**

It is a classification problem, with categorical variables. Need to classify each member belongs to which party based on voting patterns, So we build classification models on given data set

1. Naïve Bayes
2. Logistic Regression
3. Random Forest
4. Decision trees

As per business problem we no need to find number of occurrences so I am choosing error metric as “Accuracy” (True positive and True negative)

**Naïve Bayes model:**

Naive Bayes classifiers are a family of simple Probabilistic classifier based on applying Bayes’ Theorem with strong (naive) [independence](https://en.wikipedia.org/wiki/Statistical_independence) assumptions between the features.

Naïve Bayes takes and probabilities of Independent events and conditional Probabilities of Independent features

Naïve Bayes estimates a joint probability from the training data. Hence this is a Generative model

Naïve Bayes classifier predicted accurate results

**Rcode:**

>nb\_model\_congress\_data<- naiveBayes(classname~.,data = train\_congress\_data)

A-priori probabilities:

Y

democrat republican

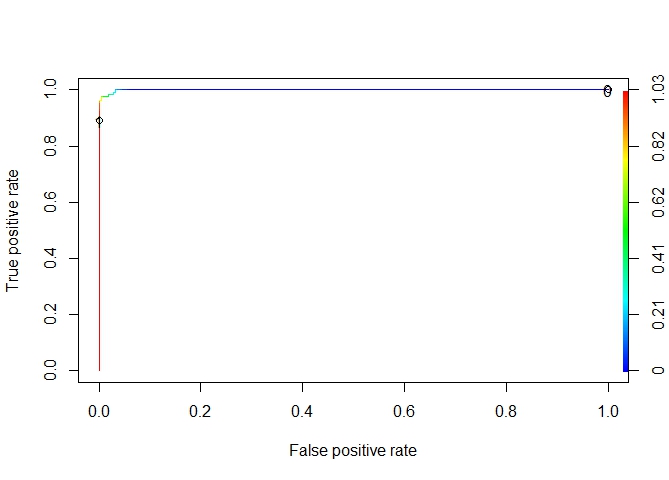
0.611898 0.388102

|  |
| --- |
| > mmetric(train\_congress\_data$classname,pred\_train\_congress\_data,  c("ACC","PRECISION","TPR","F1"))  ACC PRECISION1 PRECISION2 TPR1 TPR2 F11 F12  90.93484 96.00000 84.31373 88.88889 94.16058 92.30769 88.96552  > mmetric(test\_congress\_data$classname,pred\_test\_congress\_data,  c("ACC","PRECISION","TPR","F1"))  ACC PRECISION1 PRECISION2 TPR1 TPR2 F11 F12  96.34146 96.15385 96.66667 98.03922 93.54839 97.08738 95.08197 |
|  |
| |  | | --- | |  | |

|  |  |  |
| --- | --- | --- |
| **NAÏVE BAYES** | Error Metric | |
|  | **TRAIN** | **TEST** |
| **ACCURACY** | 90.34 | 96.34 |
| **RECALL** | 88.88 | 98.03 |
| **PRECISION** | 96.01 | 96.15 |
| **F1** | 92.30 | 97.08 |

**Logistic** **Model**:

* Logistic model estimates the probability(y/x) directly from the training data by minimizing error
* we plot ROC curve which represents True positive rate in y-axis and False positive rate in x-axis
* Her we can see threshold point at 0.9



These are the Error metric we got

|  |  |  |
| --- | --- | --- |
| **Logistic Model** | Error Metric | |
|  | **TRAIN** | **TEST** |
| **ACCURACY** | 99.43 | 93.75 |
| **RECALL** | 99.53 | 92.45 |
| **PRECISION** | 95.54 |  |
| **F1** | 92.78 | 95.84 |

**Random forest model**

Random forest is like bootstrapping algorithm with Decision tree (CART) model

Random forest tries to build multiple CART model with different sample and different initial variables

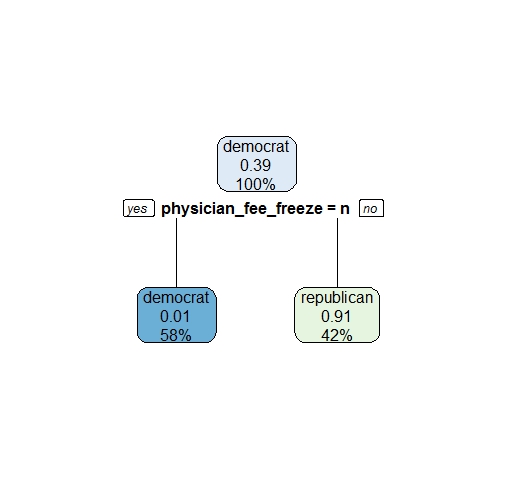
|  |
| --- |
| Modelforest <- randomForest(classname~.,data = train\_congress\_data)  > Modelforest  Call:  randomForest(formula = classname ~ ., data = train\_congress\_data)  Type of random forest: classification  Number of trees: 500  No. of variables tried at each split: 4  OOB estimate of error rate: 3.12%  Confusion matrix:  democrat republican class.error  democrat 208 8 0.03703704  republican 3 134 0.02189781 |
|  |
| |  | | --- | |  | |

Random Forests is an ensemble classifier which uses many decision tree models to predict the result

|  |  |  |
| --- | --- | --- |
| **Random Forest** | Error Metric | |
|  | **TRAIN** | **TEST** |
| **ACCURACY** | 99.43 | 97.56 |
| **RECALL** | 99.53 | 100 |
| **PRECISION** | 99.53 |  |
| **F1** | 99.53 | 98.07 |

**Decision Tree:**

We build CART model because it is a classification problem



Variable importance

physician\_fee\_freeze adoption\_of\_the\_budget\_resolution

23 16

el\_salvador\_aid education\_spending

16 16

aid\_to\_nicaraguan\_contras mx\_missile

15 14

model\_rpart <-rpart(classname~.,data=train\_congress\_data)

Node number 1: 353 observations, complexity param=0.8905109

predicted class=democrat expected loss=0.388102 P(node) =1

class counts: 216 137

probabilities: 0.612 0.388

left son=2 (205 obs) right son=3 (148 obs)

Primary splits:

physician\_fee\_freeze splits as LR, improve=139.98290, (0 missing)

adoption\_of\_the\_budget\_resolution splits as RL, improve= 95.34277, (0 missing)

education\_spending splits as LR, improve= 91.74097, (0 missing)

el\_salvador\_aid splits as LR, improve= 85.81687, (0 missing)

aid\_to\_nicaraguan\_contras splits as RL, improve= 72.12385, (0 missing)

Surrogate splits:

adoption\_of\_the\_budget\_resolution splits as RL, agree=0.873, adj=0.696, (0 split)

el\_salvador\_aid splits as LR, agree=0.873, adj=0.696, (0 split)

education\_spending splits as LR, agree=0.870, adj=0.689, (0 split)

aid\_to\_nicaraguan\_contras splits as RL, agree=0.850, adj=0.642, (0 split)

mx\_missile splits as RL, agree=0.824, adj=0.581, (0 split)

Node number 2: 205 observations

predicted class=democrat expected loss=0.009756098 P(node) =0.5807365

class counts: 203 2

probabilities: 0.990 0.010

Node number 3: 148 observations

predicted class=republican expected loss=0.08783784 P(node) =0.4192635

class counts: 13 135

probabilities: 0.088 0.912

|  |  |  |
| --- | --- | --- |
| **RPART** | Error Metric | |
|  | **TRAIN** | **TEST** |
| **ACCURACY** | 95.75 | 98.78 |
| **RECALL** | 93.98 | 93.98 |
| **PRECISION** | 99.24 |  |
| **F1** | 94.73 | 99 |

**Conclusion:** summary of the project is a classification problem, we understood the business problem , Imputed missing values, did preprocessing, divided train and test data, build four classification models we predicted the results with good accuracy.

Plan of action is to build model for each variable and see better results, how importance is given to each bill.

Thank You