MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

OVERVIEW

Explore, analyze and model a data set containing information on insurance data.

Explore, analyze and model a data set containing approximately 8000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET_FLAG, is a 1 or a 0. A "1" means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

OBJECTIVE

Objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car.

DATA SET

There are two data sets given. They are:

05/20/2017 10:55 PM 291,053 insurance-evaluation-data.csv
05/20/2017 10:55 PM 1,134,711 insurance_training_data.csv

We have to use training data to build the model

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	None
AGE	Age of Driver	Very young people tend to be risky. Maybe very old people also.
BLUEBOOK	Value of Vehicle	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect the payout if there is a crash
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probability of collision
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to file in the future
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to drive more safely
HOMEKIDS	# Children at Home	Unknown effect
HOME_VAL	Home Value	In theory, home owners tend to drive more responsibly
INCOME	Income	In theory, rich people tend to get into fewer crashes
JOB	Job Category	In theory, white collar jobs tend to be safer
KIDSDRIV	# Driving Children	When teenagers drive your car, you are more likely to get into crashes

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

MSTATUS	Marital Status	In theory, married people drive more safely
MVR_PTS	Motor Vehicle Record Points	If you get lots of traffic tickets, you tend to get into more crashes
OLDCLAIM	Total Claims (Past 5 Years)	If your total payout over the past five years was high, this suggests future payouts will be high
PARENT1	Single Parent	Unknown effect
RED_CAR	A Red Car	Urban legend says that red cars (especially red sports cars) are more risky. Is that true?
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are a more risky driver.
SEX	Gender	Urban legend says that women have less crashes then men. Is that true?
TIF	Time in Force	People who have been customers for a long time are usually more safe.
TRAVTIME	Distance to Work	Long drives to work usually suggest greater risk
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

	This color marked are two response variables.

SAMPLE DATA LOADED IN EXCEL

INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL
1	0	0	0	60	0	11	$67,349	No	$0
2	0	0	0	43	0	11	$91,449	No	$257,252
4	0	0	0	35	1	10	$16,039	No	$124,191
5	0	0	0	51	0	14		No	$306,251
6	0	0	0	50	0		$114,986	No	$243,925
7	1	2946	0	34	1	12	$125,301	Yes	$0
8	0	0	0	54	0		$18,755	No	
11	1	4021	1	37	2		$107,961	No	$333,680
12	1	2501	0	34	0	10	$62,978	No	$0
13	0	0	0	50	0	7	$106,952	No	$0
14	1	6077	0	53	0	14	$77,100	No	$0
15	0	0	0	43	0	5	$52,642	No	$209,970
16	0	0	0	55	0	11	$59,162	No	$180,232

DATA EXPLORATION

Various exploration of data set has been done through R Markdown and the program and the output has been attached

a) Summary of Statistics
b) Correlations of the data
c) Number of Rows and Columns
d) Structure of the data set

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

e) List all the variables of my data set
f) Statistical description of the data using additional packages install.packages("pastecs")

Sample from the output is given below:

```
head(trgData)
```

```
##   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV AGE HOMEKIDS YOJ   INCOME PARENT1
## 1     1           0          0        0  60        0  11  $67,349      No
## 2     2           0          0        0  43        0  11  $91,449      No
## 3     4           0          0        0  35        1  10  $16,039      No
## 4     5           0          0        0  51        0  14               No
## 5     6           0          0        0  50        0  NA $114,986      No
## 6     7           1       2946        0  34        1  12 $125,301     Yes
##   HOME_VAL MSTATUS SEX     EDUCATION          JOB TRAVTIME    CAR_USE
## 1       $0    z_No   M           PhD  Professional       14    Private
## 2 $257,252    z_No   M z_High School z_Blue Collar       22 Commercial
## 3 $124,191     Yes z_F z_High School      Clerical        5    Private
## 4 $306,251     Yes   M  <High School z_Blue Collar       32    Private
## 5 $243,925     Yes z_F           PhD        Doctor       36    Private
## 6       $0    z_No z_F     Bachelors z_Blue Collar       46 Commercial
##   BLUEBOOK TIF   CAR_TYPE RED_CAR OLDCLAIM CLM_FREQ REVOKED MVR_PTS
## 1  $14,230  11    Minivan     yes   $4,461        2      No       3
## 2  $14,940   1    Minivan     yes       $0        0      No       0
## 3   $4,010   4      z_SUV      no  $38,690        2      No       3
## 4  $15,440   7    Minivan     yes       $0        0      No       0
## 5  $18,000   1      z_SUV      no  $19,217        2     Yes       3
## 6  $17,430   1 Sports Car      no       $0        0      No       0
##   CAR_AGE          URBANICITY
## 1      18 Highly Urban/ Urban
## 2       1 Highly Urban/ Urban
## 3      10 Highly Urban/ Urban
## 4       6 Highly Urban/ Urban
## 5      17 Highly Urban/ Urban
## 6       7 Highly Urban/ Urban
```

```
summary(trgData)
```

```
##      INDEX        TARGET_FLAG       TARGET_AMT        KIDSDRIV
##  Min.   :    1   Min.   :0.0000   Min.   :     0   Min.   :0.0000
##  1st Qu.: 2559   1st Qu.:0.0000   1st Qu.:     0   1st Qu.:0.0000
##  Median : 5133   Median :0.0000   Median :     0   Median :0.0000
##  Mean   : 5152   Mean   :0.2638   Mean   :  1504   Mean   :0.1711
##  3rd Qu.: 7745   3rd Qu.:1.0000   3rd Qu.:  1036   3rd Qu.:0.0000
##  Max.   :10302   Max.   :1.0000   Max.   :107586   Max.   :4.0000
##
##       AGE           HOMEKIDS           YOJ            INCOME
```

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 - HOMEWORK 3 WEEK8

```
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   $0      : 615
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0           : 445
##  Median :45.00   Median :0.0000   Median :11.0   $26,840 :   4
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5   $48,509 :   4
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0   $61,790 :   4
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0   $107,375:   3
##  NA's   :6                        NA's   :454    (Other) :7086
##  PARENT1        HOME_VAL    MSTATUS      SEX              EDUCATION
##  No :7084   $0      :2294   Yes :4894   M  :3786   <High School :1203
##  Yes:1077           : 464   z_No:3267   z_F:4375   Bachelors    :2242
##             $111,129:   3                          Masters      :1658
##             $115,249:   3                          PhD          : 728
##             $123,109:   3                          z_High School:2330
##             $153,061:   3
##             (Other) :5391
##             JOB          TRAVTIME          CAR_USE        BLUEBOOK
##  z_Blue Collar:1825   Min.   :  5.00   Commercial:3029   $1,500 : 157
##  Clerical     :1271   1st Qu.: 22.00   Private   :5132   $6,000 :  34
##  Professional :1117   Median : 33.00                     $5,800 :  33
##  Manager      : 988   Mean   : 33.49                     $6,200 :  33
##  Lawyer       : 835   3rd Qu.: 44.00                     $6,400 :  31
##  Student      : 712   Max.   :142.00                     $5,900 :  30
##  (Other)      :1413                                      (Other):7843
##       TIF               CAR_TYPE    RED_CAR       OLDCLAIM
##  Min.   : 1.000   Minivan    :2145   no :5783   $0     :5009
##  1st Qu.: 1.000   Panel Truck: 676   yes:2378   $1,310 :   4
##  Median : 4.000   Pickup     :1389              $1,391 :   4
##  Mean   : 5.351   Sports Car : 907              $4,263 :   4
##  3rd Qu.: 7.000   Van        : 750              $1,105 :   3
##  Max.   :25.000   z_SUV      :2294              $1,332 :   3
##                                                 (Other):3134

##                                                 (Other):3134
##     CLM_FREQ       REVOKED        MVR_PTS          CAR_AGE
##  Min.   :0.0000   No :7161   Min.   : 0.000   Min.   :-3.000
##  1st Qu.:0.0000   Yes:1000   1st Qu.: 0.000   1st Qu.: 1.000
##  Median :0.0000              Median : 1.000   Median : 8.000
##  Mean   :0.7986              Mean   : 1.696   Mean   : 8.328
##  3rd Qu.:2.0000              3rd Qu.: 3.000   3rd Qu.:12.000
##  Max.   :5.0000              Max.   :13.000   Max.   :28.000
##                                               NA's   :510
##                 URBANICITY
##  Highly Urban/ Urban  :6492
##  z_Highly Rural/ Rural:1669
##
##
##
##
##
```

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

DATA PREPARATION

1. In order to get the analysis, We have installed funModelling package for this. (viz. install.packages("funModeling")). This gives detailed report of about missing any data. – Please refer the output of RMARKDOWN file.

```
df_status(trgData)
```

```
##       variable q_zeros p_zeros q_na p_na q_inf p_inf    type unique
## 1        INDEX       0    0.00    0 0.00     0     0 integer   8161
## 2  TARGET_FLAG    6008   73.62    0 0.00     0     0 integer      2
## 3   TARGET_AMT    6008   73.62    0 0.00     0     0 numeric   1949
## 4     KIDSDRIV    7180   87.98    0 0.00     0     0 integer      5
## 5          AGE       0    0.00    6 0.07     0     0 integer     60
## 6     HOMEKIDS    5289   64.81    0 0.00     0     0 integer      6
## 7          YOJ     625    7.66  454 5.56     0     0 integer     21
## 8       INCOME       0    0.00    0 0.00     0     0  factor   6613
## 9      PARENT1       0    0.00    0 0.00     0     0  factor      2
## 10    HOME_VAL       0    0.00    0 0.00     0     0  factor   5107
## 11     MSTATUS       0    0.00    0 0.00     0     0  factor      2
## 12         SEX       0    0.00    0 0.00     0     0  factor      2
## 13   EDUCATION       0    0.00    0 0.00     0     0  factor      5
## 14         JOB       0    0.00    0 0.00     0     0  factor      9
## 15    TRAVTIME       0    0.00    0 0.00     0     0 integer     97
```

```
## 16     CAR_USE       0    0.00    0 0.00     0     0  factor      2
## 17    BLUEBOOK       0    0.00    0 0.00     0     0  factor   2789
## 18         TIF       0    0.00    0 0.00     0     0 integer     23
## 19    CAR_TYPE       0    0.00    0 0.00     0     0  factor      6
## 20     RED_CAR       0    0.00    0 0.00     0     0  factor      2
## 21    OLDCLAIM       0    0.00    0 0.00     0     0  factor   2857
## 22    CLM_FREQ    5009   61.38    0 0.00     0     0 integer      6
## 23     REVOKED       0    0.00    0 0.00     0     0  factor      2
## 24     MVR_PTS    3712   45.48    0 0.00     0     0 integer     13
## 25     CAR_AGE       3    0.04  510 6.25     0     0 integer     30
## 26  URBANICITY       0    0.00    0 0.00     0     0  factor      2
```

```
######## checking whether any cell has NA or Infinite
apply(trgData, 2, function(x) any(is.na(x)))
```

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##       FALSE       FALSE       FALSE       FALSE        TRUE       FALSE
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##        TRUE       FALSE       FALSE       FALSE       FALSE       FALSE
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##       FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##       FALSE       FALSE       FALSE       FALSE       FALSE       FALSE
##     CAR_AGE  URBANICITY
##        TRUE       FALSE
```

2. Put the Data into buckets

Individual columns or combination of any columns can be separated out and put them into a smaller buckets for sample analysis

Sample of the output is given below

```
bucket.AGE<-trgData['AGE']
summary(bucket.AGE)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   16.00   39.00   45.00   44.79   51.00   81.00       6
```

```
#bucket.indus

bucket.EDUCATION.JOB<-cbind(trgData$EDUCATION,trgData$JOB)
summary(bucket.EDUCATION.JOB)
```

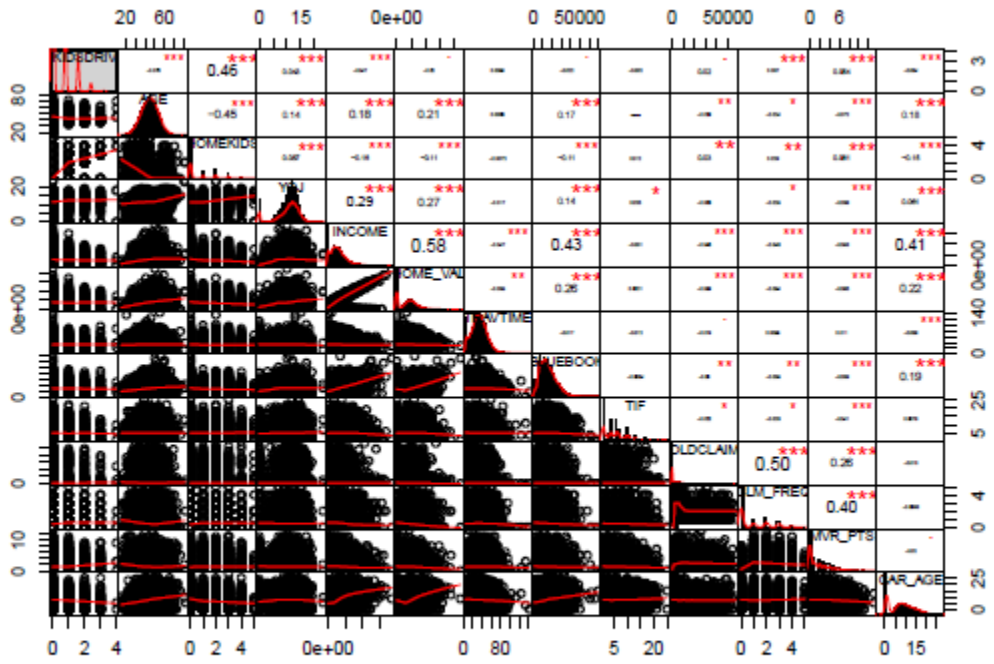
```
##        V1              V2
##  Min.   :1.000   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:3.000
##  Median :3.000   Median :6.000
##  Mean   :3.091   Mean   :5.687
##  3rd Qu.:5.000   3rd Qu.:8.000
##  Max.   :5.000   Max.   :9.000
```

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

DATA VISUALIZATION

Sample data visualization has been given in RMARKDOWN output.

Installed Performance Analytics Package and found How Data is Correlated each other? – Please see below. Also refer the output of the RMARKDOWN FILE



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

The following data requires cleanup

Income – Need to remove characters like $
HOME_VAL – need to remove characters like $
MSTATUS : remove z_
SEX – remove z_
EDUCATION – remove z_
JOB – remove prefix z_
BLUEBOOK remove character like $
CAR_TYPE – remove prefix z_
OLDCLAIM – remove character like $

MUTHUKUMAR SRINIVASAN & RAJAGOPAL SRINIVASAN
DATA 621 – HOMEWORK 3 WEEK8

BUILD MODEL & SELECT MODEL

Please see the output of the markdown file.

CONCLUSION

Rmarkdown file

Output pdf file

Data file

This word document converted to PDF file

Are all available in https://github.com/muthukumars/DATA-621/tree/master/Week10-Homework4

THANK YOU