HW4

Business Analytics and Data Mining

Zachary Palmore

4/17/2021

Assignment 4

```r
# Packages
library(tidyverse)
library(kableExtra)
```

Purpose

In this homework assignment, we will 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.

Our 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. We can only use the variables given (or variables derived from the variables provided). Below is a short description of the variables of interest in the data set:

```r
# short descriptions of variables as table from matrix
vardesc <- data.frame(matrix(c(
'INDEX',    'Identification variable',
'TARGET_FLAG',  'Was car in a crash? 1 = Yes, 0 = No',
'TARGET_AMT',   'Cost of car crash',
'AGE',  'Age of driver',
'BLUEBOOK', 'Value of vehicle',
'CAR_AGE',  'Vehicle age',
'CAR_TYPE', 'Type of car',
'CAR_USE', 'Main purpose the vehicle is used for',
'CLM_FREQ', 'Number of claims filed in past five years',
'EDUCATION',    'Maximum education level',
'HOMEKIDS', 'Number of children at home',
'HOME_VAL', 'Value of driver\'s home',
'INCOME',   'Annual income of the driver',
'JOB',  'Type of job by standard collar categories',
'KIDSDRIV', 'Number of children who drive',
'MSTATUS',  'Marital status',
'MVR_PTS',  'Motor vehicle inspection points',
'OLDCLAIM', 'Total claims payout in past five years',
'PARENT1',  'Single parent status',
'RED_CAR',  '1 if car is red, 0 if not',
'REVOKED',  'License revoked in past 7 years status',
'SEX',  'Driver gender',
'TIF',  'Time in force',
'TRAVETIME',    'Distance to work in minutes',
'URBANICITY',   'Category of how urban the area the driver lives is',
'YOJ', 'Number of years on the job'
),  byrow = TRUE, ncol = 2))
colnames(vardesc) <- c('Variable', 'Description')
kbl(vardesc, booktabs = T, caption = "Variable Descriptions") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"), full_width = F)
```

Table 1: Variable Descriptions

Variable	Description
INDEX	Identification variable
TARGET_FLAG	Was car in a crash? 1 = Yes, 0 = No
TARGET_AMT	Cost of car crash
AGE	Age of driver
BLUEBOOK	Value of vehicle
CAR_AGE	Vehicle age
CAR_TYPE	Type of car
CAR_USE	Main purpose the vehicle is used for
CLM_FREQ	Number of claims filed in past five years
EDUCATION	Maximum education level
HOMEKIDS	Number of children at home
HOME_VAL	Value of driver's home
INCOME	Annual income of the driver
JOB	Type of job by standard collar categories
KIDSDRIV	Number of children who drive
MSTATUS	Marital status
MVR_PTS	Motor vehicle inspection points
OLDCLAIM	Total claims payout in past five years
PARENT1	Single parent status
RED_CAR	1 if car is red, 0 if not
REVOKED	License revoked in past 7 years status
SEX	Driver gender
TIF	Time in force
TRAVETIME	Distance to work in minutes
URBANICITY	Category of how urban the area the driver lives is
YOJ	Number of years on the job

Introduction

```r
tdata <- read.csv(
  "https://raw.githubusercontent.com/palmorezm/msds/main/621/HW4/insurance_training_data.csv")
edata <- read.csv(
  "https://raw.githubusercontent.com/palmorezm/msds/main/621/HW4/insurance-evaluation-data.csv")
```

```r
initialobs <- tdata[1:4,]
kbl(t(initialobs), booktabs = T, caption = "Initial Observations") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"), full_width = F) %>%
  add_header_above(c(" ", " ", "Row Number", " ", " ")) %>%
  footnote(c("Includes the first four observations of all variables in the data"))
```

Table 2: Initial Observations

		Row Number		
	1	2	3	4
INDEX	1	2	4	5
TARGET_FLAG	0	0	0	0
TARGET_AMT	0	0	0	0
KIDSDRIV	0	0	0	0
AGE	60	43	35	51
HOMEKIDS	0	0	1	0
YOJ	11	11	10	14
INCOME	$67,349	$91,449	$16,039	
PARENT1	No	No	No	No
HOME_VAL	$0	$257,252	$124,191	$306,251
MSTATUS	z_No	z_No	Yes	Yes
SEX	M	M	z_F	M
EDUCATION	PhD	z_High School	z_High School	<High School
JOB	Professional	z_Blue Collar	Clerical	z_Blue Collar
TRAVTIME	14	22	5	32
CAR_USE	Private	Commercial	Private	Private
BLUEBOOK	$14,230	$14,940	$4,010	$15,440
TIF	11	1	4	7
CAR_TYPE	Minivan	Minivan	z_SUV	Minivan
RED_CAR	yes	yes	no	yes
OLDCLAIM	$4,461	$0	$38,690	$0
CLM_FREQ	2	0	2	0
REVOKED	No	No	No	No
MVR_PTS	3	0	3	0
CAR_AGE	18	1	10	6
URBANICITY	Highly Urban/ Urban	Highly Urban/ Urban	Highly Urban/ Urban	Highly Urban/ Urban

Note:
Includes the first four observations of all variables in the data

Data Exploration

Describe the size and the variables in the insurance training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job. Some suggestions are given below. Please do NOT treat this as a check list of things to do to complete the assignment. You should have your own thoughts on what to tell the boss. These are just ideas. a. Mean / Standard Deviation / Median b. Bar Chart or Box Plot of the data c. Is the data correlated to the target variable (or to other variables?) d. Are any of the variables missing and need to be imputed "fixed"?

```
# theoretical effects
vareffects <- data.frame(matrix(c(
'INDEX',    'None',
'TARGET_FLAG',  'None',
'TARGET_AMT',   'None',
'AGE',  'Youngest and Oldest may have higher risk of accident',
'BLUEBOOK', 'Unknown on probability of collision but correlated with payout',
'CAR_AGE',  'Unknown on probability of collision but correlated with payout',
'CAR_TYPE', 'Unknown on probability of collision but correlated with payout',
'CAR_USE', 'Commerical vehicles might increase risk of accident',
'CLM_FREQ', 'Higher claim frequency increases likelihood of future claims',
'EDUCATION',    'Theoretically higher education levels lower risk',
'HOMEKIDS', 'Unknown',
'HOME_VAL', 'Theoretically home owners reduce risk due to more responsible driving',
'INCOME',   'Theoretically wealthier drivers have fewer accidents',
'JOB',  'Theoretically white collar+ jobs are safer',
'KIDSDRIV', 'Increased risk of accident from inexperienced driver',
'MSTATUS',  'Theoretically married people drive safer',
'MVR_PTS',  'Increased risk of accident',
'OLDCLAIM', 'Increased risk of higher payout with previous payout',
'PARENT1',  'Unknown',
'RED_CAR',  'Theoretically increased risk of accident based on urban legend',
'REVOKED',  'Increased risk of accident if revoked',
'SEX',  'Theoretically increased risk of accident for women based on urban legend',
'TIF',  'Decreased risk for those who have greater loyalty',
'TRAVETIME',    'Longer distances increase risk of accident',
'URBANICITY',   'The more urban the area the greater the risk of accident',
'YOJ', 'Decreased risk for those with greater longevity'
),  byrow = TRUE, ncol = 2))
colnames(vareffects) <- c('Variable', 'Effect')
kbl(vareffects, booktabs = T, caption = "Theoretical Variable Effects") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"), full_width = F)
```

Table 3: Theoretical Variable Effects

Variable	Effect
INDEX	None
TARGET_FLAG	None
TARGET_AMT	None
AGE	Youngest and Oldest may have higher risk of accident
BLUEBOOK	Unknown on probability of collision but correlated with payout
CAR_AGE	Unknown on probability of collision but correlated with payout
CAR_TYPE	Unknown on probability of collision but correlated with payout
CAR_USE	Commerical vehicles might increase risk of accident
CLM_FREQ	Higher claim frequency increases likelihood of future claims
EDUCATION	Theoretically higher education levels lower risk
HOMEKIDS	Unknown
HOME_VAL	Theoretically home owners reduce risk due to more responsible driving
INCOME	Theoretically wealthier drivers have fewer accidents
JOB	Theoretically white collar+ jobs are safer
KIDSDRIV	Increased risk of accident from inexperienced driver
MSTATUS	Theoretically married people drive safer
MVR_PTS	Increased risk of accident
OLDCLAIM	Increased risk of higher payout with previous payout
PARENT1	Unknown
RED_CAR	Theoretically increased risk of accident based on urban legend
REVOKED	Increased risk of accident if revoked
SEX	Theoretically increased risk of accident for women based on urban legend
TIF	Decreased risk for those who have greater loyalty
TRAVETIME	Longer distances increase risk of accident
URBANICITY	The more urban the area the greater the risk of accident
YOJ	Decreased risk for those with greater longevity

```
tdata.nas <- lapply(tdata, function(x) sum(is.na(x)))
tdata.types <- lapply(tdata, function(x) class(x))
tdata.firstob <- lapply(tdata, function(x) head(x, 1))
tdata.uniques <- lapply(tdata, function(x) length(unique(factor(x))))
tdata.tbl.natypes <- cbind(tdata.nas, tdata.types, tdata.firstob, tdata.uniques)
colnames(tdata.tbl.natypes) <- c("Quantity Missing", "Data Type", "First Observation", "Unique Factors")
kbl(tdata.tbl.natypes, booktabs = T, caption = "Data Characteristics") %>%
  kable_styling(latex_options = c("striped", "HOLD_position"), full_width = F)
```

Table 4: Data Characteristics

	Quantity Missing	Data Type	First Observation	Unique Factors
INDEX	0	integer	1	8161
TARGET_FLAG	0	integer	0	2
TARGET_AMT	0	numeric	0	1949
KIDSDRIV	0	integer	0	5
AGE	6	integer	60	61
HOMEKIDS	0	integer	0	6
YOJ	454	integer	11	22
INCOME	0	character	$67,349	6613
PARENT1	0	character	No	2
HOME_VAL	0	character	$0	5107
MSTATUS	0	character	z_No	2
SEX	0	character	M	2
EDUCATION	0	character	PhD	5
JOB	0	character	Professional	9
TRAVTIME	0	integer	14	97
CAR_USE	0	character	Private	2
BLUEBOOK	0	character	$14,230	2789
TIF	0	integer	11	23
CAR_TYPE	0	character	Minivan	6
RED_CAR	0	character	yes	2
OLDCLAIM	0	character	$4,461	2857
CLM_FREQ	0	integer	2	6
REVOKED	0	character	No	2
MVR_PTS	0	integer	3	13
CAR_AGE	510	integer	18	31
URBANICITY	0	character	Highly Urban/ Urban	2

```r
tdata.summary.tbl <- summary(tdata)
kbl(t(tdata.summary.tbl), booktabs = T, caption = "Data Characteristics") %>%
  kable_styling(latex_options = c("striped", "scale_down", "hold_position"), full_width = F)
```

Table 5: Data Characteristics

INDEX	Min. : 1	1st Qu.: 2559	Median : 5133	Mean : 5152	3rd Qu.: 7745	Max. :10302	NA
TARGET_FLAG	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.2638	3rd Qu.:1.0000	Max. :1.0000	NA
TARGET_AMT	Min. : 0	1st Qu.: 0	Median : 0	Mean : 1504	3rd Qu.: 1036	Max. :107586	NA
KIDSDRIV	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.1711	3rd Qu.:0.0000	Max. :4.0000	NA
AGE	Min. :16.00	1st Qu.:39.00	Median :45.00	Mean :44.79	3rd Qu.:51.00	Max. :81.00	NA's :6
HOMEKIDS	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.7212	3rd Qu.:1.0000	Max. :5.0000	NA
YOJ	Min. : 0.0	1st Qu.: 9.0	Median :11.0	Mean :10.5	3rd Qu.:13.0	Max. :23.0	NA's :454
INCOME	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
PARENT1	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
HOME_VAL	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
MSTATUS	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
SEX	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
EDUCATION	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
JOB	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
TRAVTIME	Min. : 5.00	1st Qu.: 22.00	Median : 33.00	Mean : 33.49	3rd Qu.: 44.00	Max. :142.00	NA
CAR_USE	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
BLUEBOOK	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
TIF	Min. : 1.000	1st Qu.: 1.000	Median : 4.000	Mean : 5.351	3rd Qu.: 7.000	Max. :25.000	NA
CAR_TYPE	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
RED_CAR	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
OLDCLAIM	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
CLM_FREQ	Min. :0.0000	1st Qu.:0.0000	Median :0.0000	Mean :0.7986	3rd Qu.:2.0000	Max. :5.0000	NA
REVOKED	Length:8161	Class :character	Mode :character	NA	NA	NA	NA
MVR_PTS	Min. : 0.000	1st Qu.: 0.000	Median : 1.000	Mean : 1.696	3rd Qu.: 3.000	Max. :13.000	NA
CAR_AGE	Min. :-3.000	1st Qu.: 1.000	Median : 8.000	Mean : 8.328	3rd Qu.:12.000	Max. :28.000	NA's :510
URBANICITY	Length:8161	Class :character	Mode :character	NA	NA	NA	NA

Data Preparation

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this. Here are some possible transformations. a. Fix missing values (maybe with a Mean or Median value) b. Create flags to suggest if a variable was missing c. Transform data by putting it into buckets d. Mathematical transforms such as log or square root (or use Box-Cox) e. Combine variables (such as ratios or adding or multiplying) to create new variables

Model Building

Using the training data set, build at least two different multiple linear regression models and three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Stepwise, use a different approach such as trees, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done.

Discuss the coefficients in the models, do they make sense? For example, if a person has a lot of traffic tickets, you would reasonably expect that person to have more car crashes. If the coefficient is negative (suggesting that the person is a safer driver), then that needs to be discussed. Are you keeping the model even though it is counter intuitive? Why? The boss needs to know.

Model Selection

Decide on the criteria for selecting the best multiple linear regression model and the best binary logistic regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your models. For the multiple linear regression model, will you use a metric such as Adjusted R2 , RMSE, etc.? Be sure to explain how you can make inferences from the model, discuss multi-collinearity issues (if any), and discuss other relevant model output. Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2 , (c) F-statistic, and (d) residual plots. For the binary logistic regression model, will you use a metric such as log likelihood, AIC, ROC curve, etc.? Using the training data set, evaluate the binary logistic regression model based on (a) accuracy, (b) classification error rate, (c) precision, (d) sensitivity, (e) specificity, (f) F1 score, (g) AUC, and (h) confusion matrix. Make predictions using the evaluation data set.