HW 4

Team 2

April 24, 2019

Contents

OVERVIEW

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

Dependencies

Replication of our work requires the following packages in Rstudio:

```
library(ggplot2)
library(ggpubr)
library(dplyr)
library(tidyr)
library(corrplot)
library(randomForest)
library(olsrr)
library(psych)
```

Objective

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.

PART 1: DATA EXPLORATION

First, we read the data as a csv and then examined the below variable from the `training` dataset.

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
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

Summary Statistics

We look at summary of the data below. Note that rows marked with * indicate categorical variables that were converted to numeric.

	vars	n	mean	sd	min	max	range	se
TARGET_FLAG	1	8161	0.26	0.44	0	1.0	1.0	0.00
TARGET_AMT	2	8161	1504.32	4704.03	0	107586.1	107586.1	52.07
KIDSDRIV	3	8161	0.17	0.51	0	4.0	4.0	0.01
AGE	4	8155	44.79	8.63	16	81.0	65.0	0.10
HOMEKIDS	5	8161	0.72	1.12	0	5.0	5.0	0.01
YOJ	6	7707	10.50	4.09	0	23.0	23.0	0.05
INCOME*	7	8161	2875.55	2090.68	1	6613.0	6612.0	23.14
PARENT1*	8	8161	1.13	0.34	1	2.0	1.0	0.00
HOME_VAL*	9	8161	1684.89	1697.38	1	5107.0	5106.0	18.79
MSTATUS*	10	8161	1.40	0.49	1	2.0	1.0	0.01
SEX*	11	8161	1.54	0.50	1	2.0	1.0	0.01
EDUCATION*	12	8161	3.09	1.44	1	5.0	4.0	0.02
JOB*	13	8161	5.69	2.68	1	9.0	8.0	0.03
TRAVTIME	14	8161	33.49	15.91	5	142.0	137.0	0.18
CAR_USE*	15	8161	1.63	0.48	1	2.0	1.0	0.01
BLUEBOOK*	16	8161	1283.62	893.51	1	2789.0	2788.0	9.89
TIF	17	8161	5.35	4.15	1	25.0	24.0	0.05
CAR_TYPE*	18	8161	3.53	1.97	1	6.0	5.0	0.02
RED_CAR*	19	8161	1.29	0.45	1	2.0	1.0	0.01
OLDCLAIM*	20	8161	552.27	862.20	1	2857.0	2856.0	9.54
CLM_FREQ	21	8161	0.80	1.16	0	5.0	5.0	0.01
REVOKED*	22	8161	1.12	0.33	1	2.0	1.0	0.00
MVR_PTS	23	8161	1.70	2.15	0	13.0	13.0	0.02
CAR_AGE	24	7651	8.33	5.70	-3	28.0	31.0	0.07
URBANICITY*	25	8161	1.20	0.40	1	2.0	1.0	0.00

From this, we can see that some variables have missing values, so we will have to impute them later. For `HOME_VAL` and `INCOME` values, we imputed them as the mean, which doesn't change the variance. However, this assumes that these people have both income and homes. If not, this would bias our models.

We must also convert the currency data to a numeric form. One record has a car age of -3, which was also replaced with the mean. There are 4 records where `JOB` is blank, which will be treated as its own category.

```
for(i in c(4,6,7,9,24)){
  training[is.na(training[,i]), i] <- mean(training[,i], na.rm = TRUE)
}


training$CAR_AGE[training$CAR_AGE<0] <- mean(training$CAR_AGE)

training$CAR_AGE[training$CAR_AGE<0]
```

```
FALSE numeric(0)
```

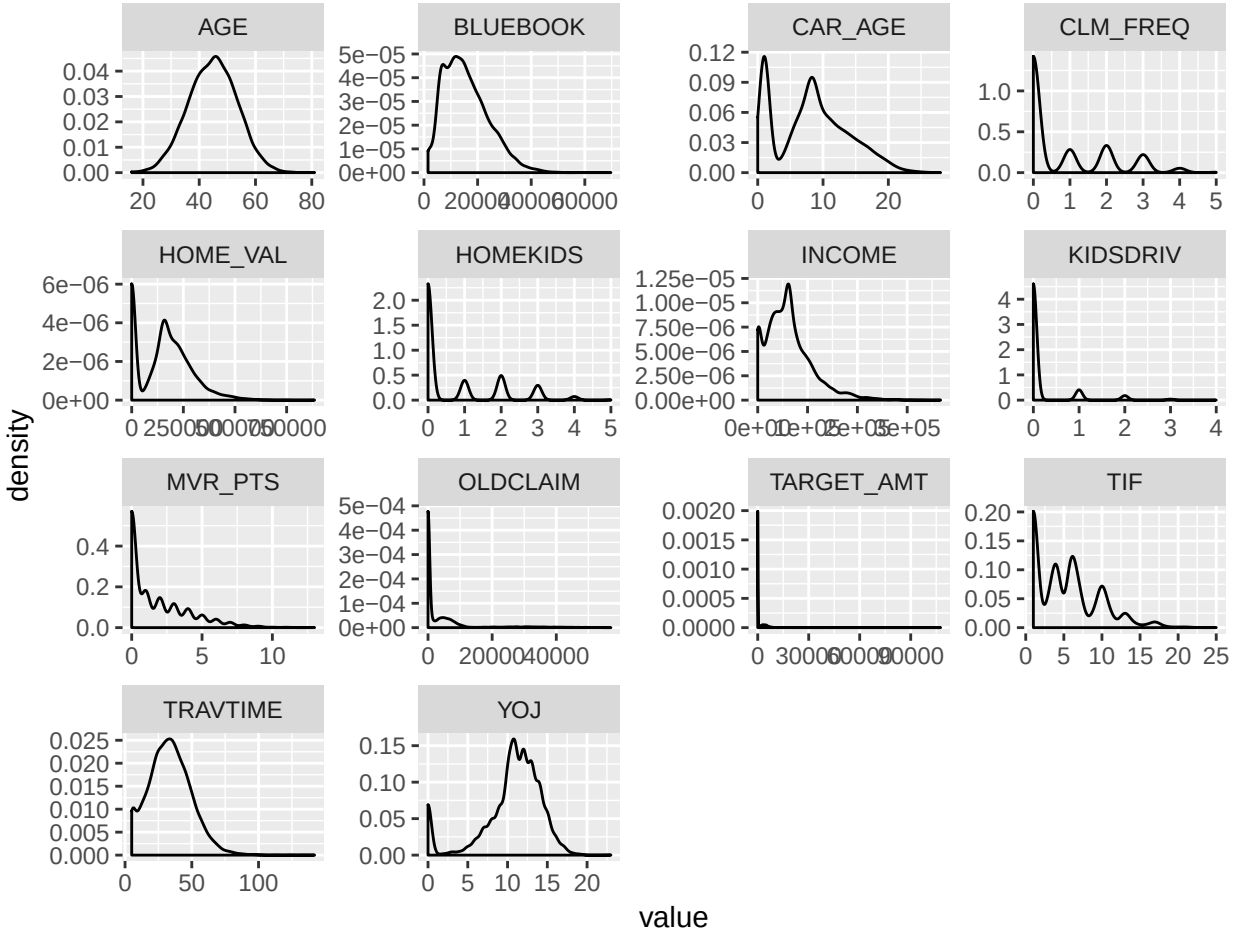
```
training$JOB[training$JOB=='']
```

```
FALSE   [1]
FALSE  [71]
FALSE [141]
FALSE [211]
FALSE [281]
FALSE [351]
```

```
FALSE [421]
FALSE [491]
FALSE 9 Levels:  Clerical Doctor Home Maker Lawyer Manager ... z_Blue Collar
```

Density

Below, we examine the distribution of values for each of the variables.

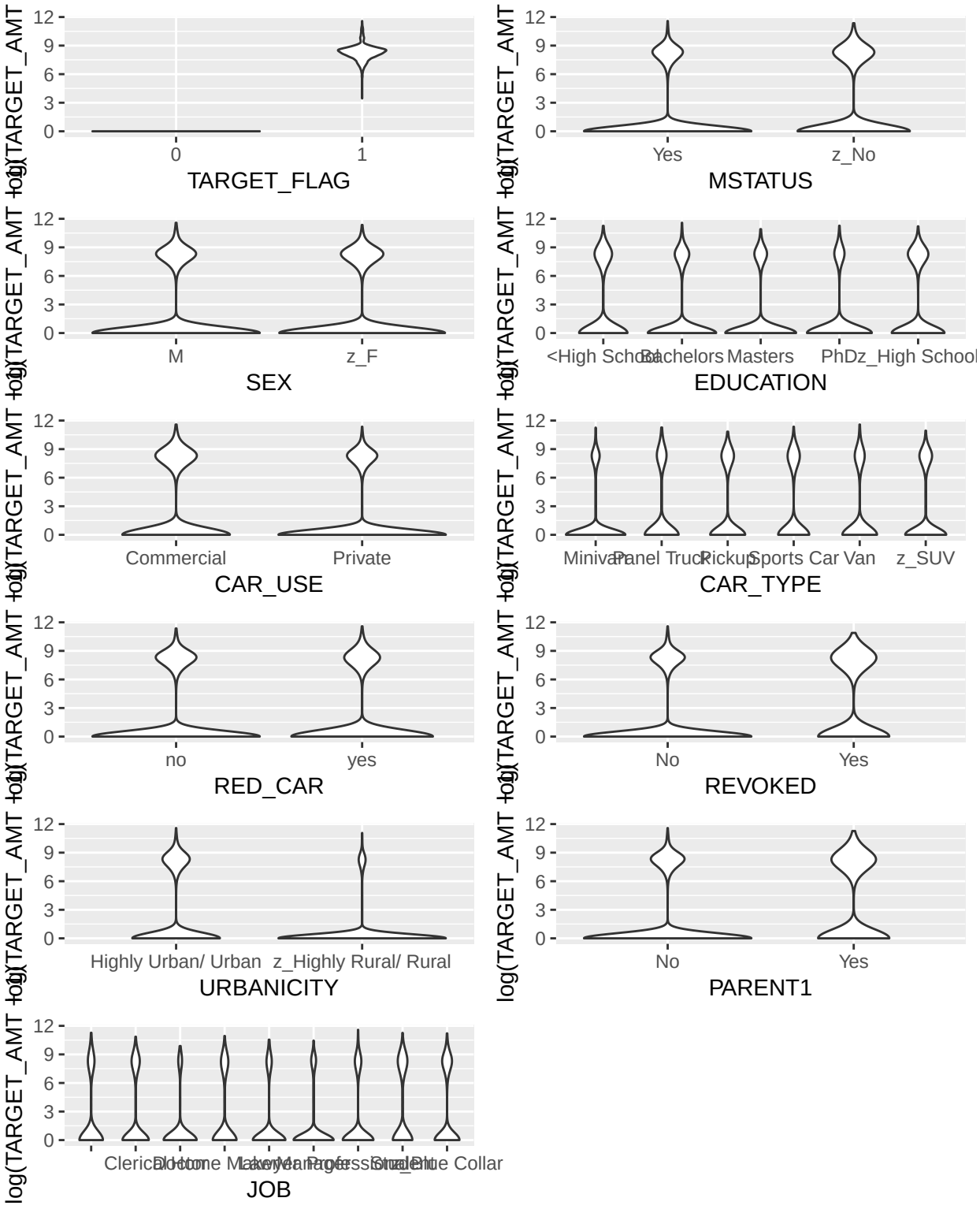


Variables, `AGE` and `YOJ`, appear to be fairly normally distributed. However, `CLM_FREQ`, `HOME_VAL`, `HOME_KIDS`, `KIDSDRIV`, and `MVR_PTS` appear to follow either a continuous or discrete quasi-Poisson process.

In any case, claim frequency, `HOME_KIDS`, `KIDSDRIV`, `MVR_PTS`, `OLDCLAIM`, `TIF`, `TRAVTIME`, and `YOJ` are count values. Log transforms will be inappropriate and skew the data, particularly the ones with

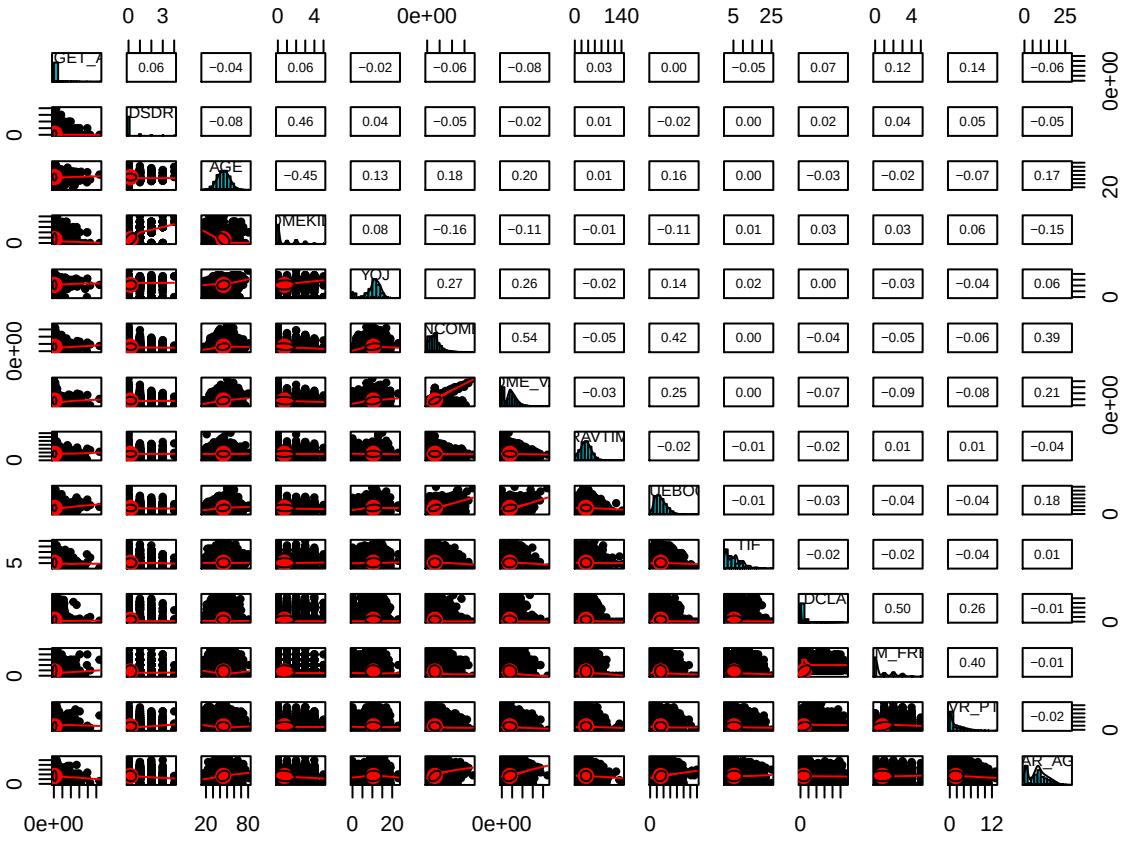
---

?????? Sentence trails off. Please complete thought - jm



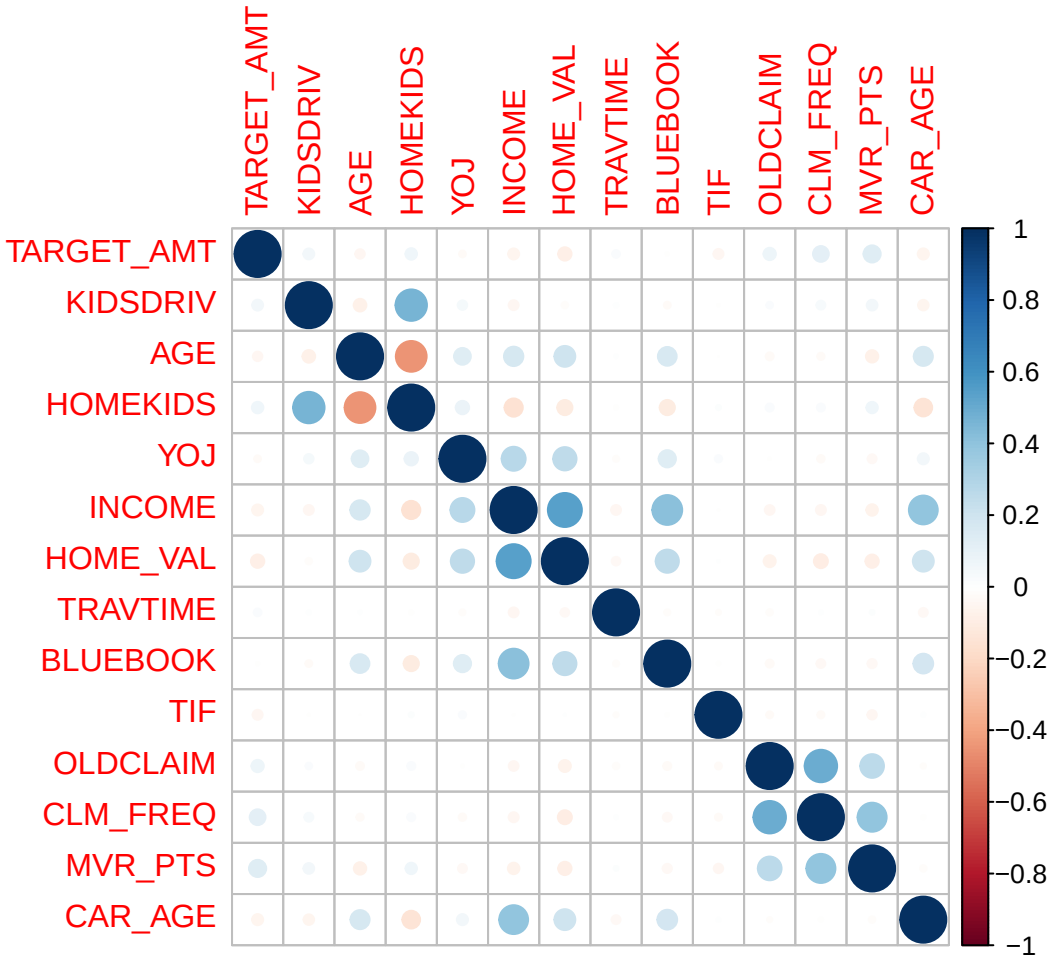
Scatter plot matrix

We then build scatter plot matrix for continious variables



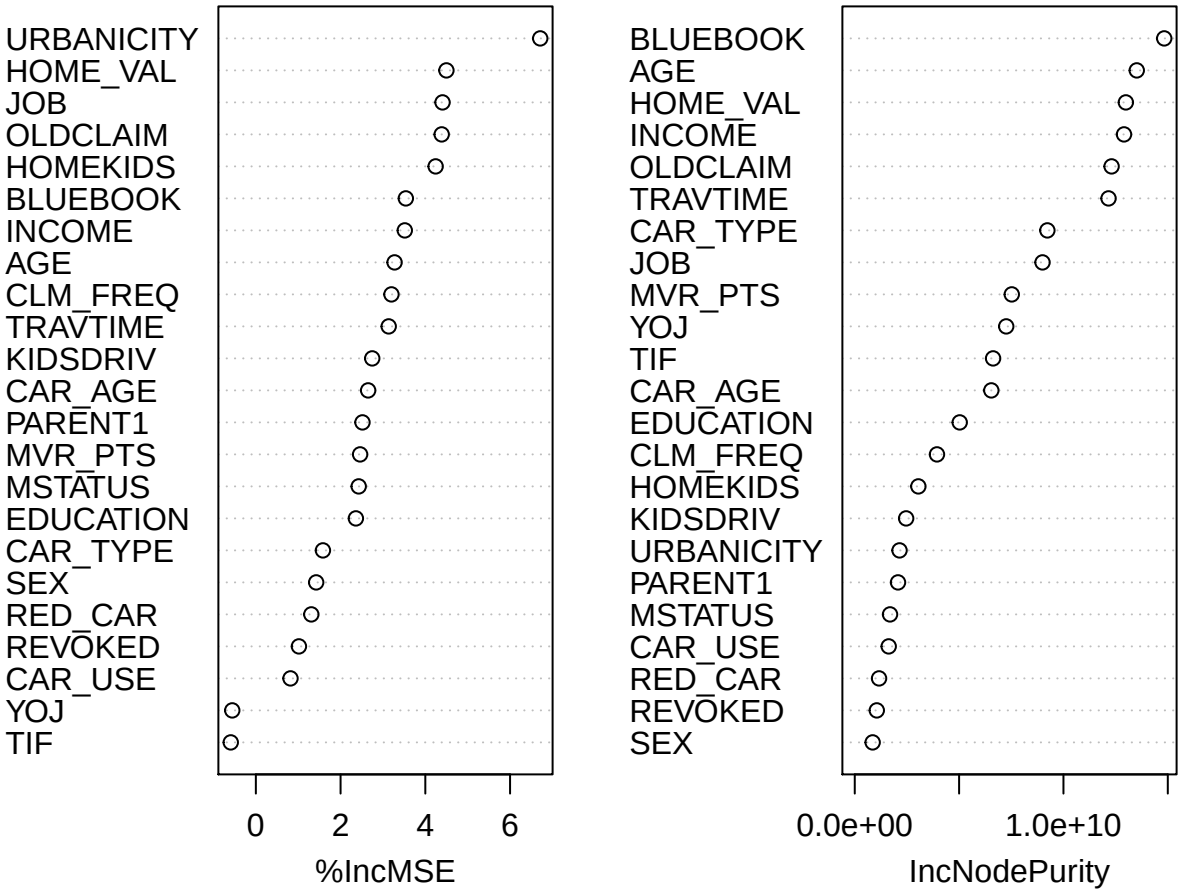
Correlation

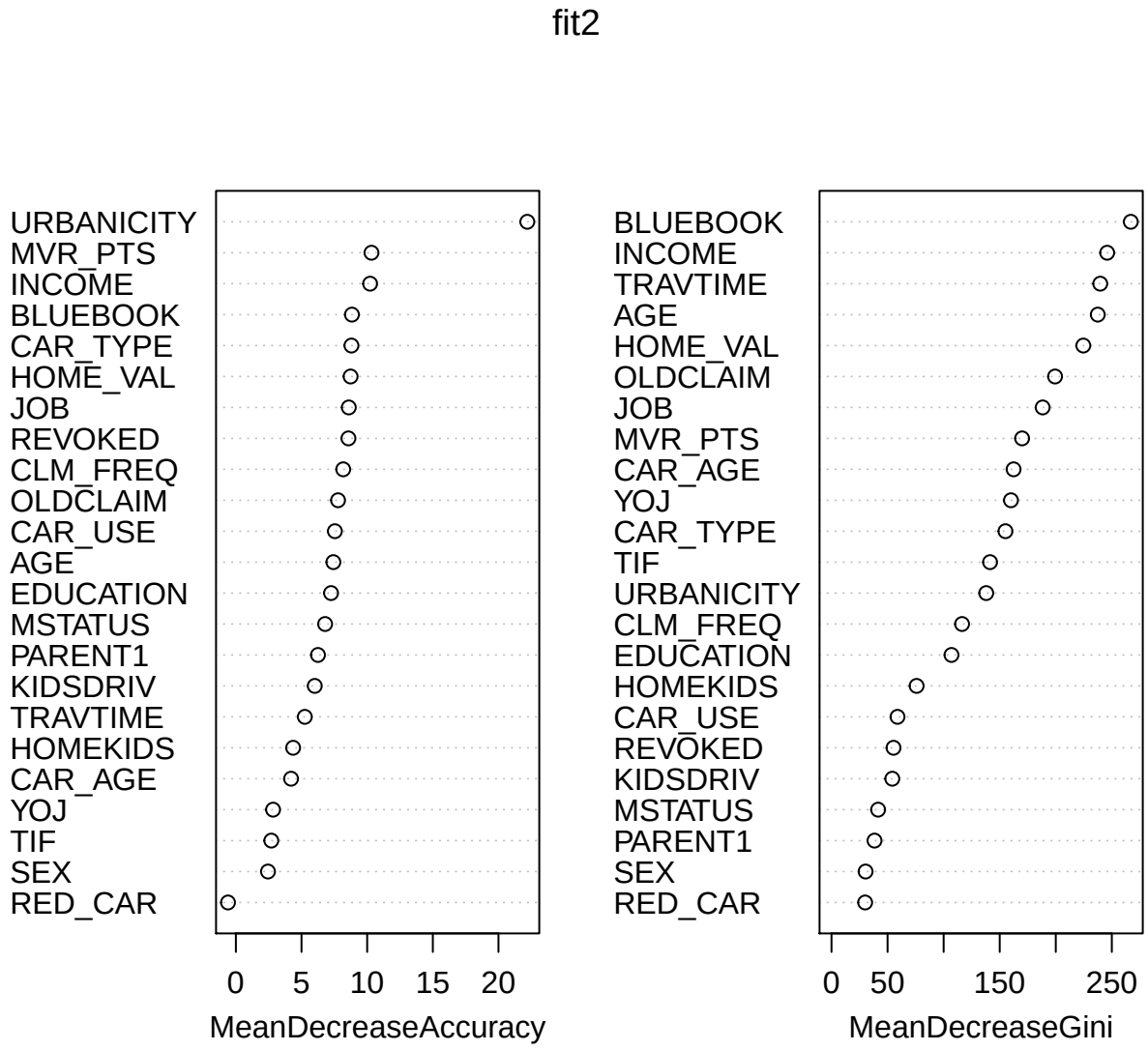
We can see our correlation matrix below. A dark blue circle represents a strong positive relationship and a dark red circle represents a strong negative relationship between two variables.



Finally, we can use the `randomforest` package to verify our assumptions from the correlation plot.

fit1


URBANICITY
HOME_VAL
JOB
OLDCLAIM
HOMEKIDS
BLUEBOOK
INCOME
AGE
CLM_FREQ
TRAVTIME
KIDSDRIV
CAR_AGE
PARENT1
MVR_PTS
MSTATUS
EDUCATION
CAR_TYPE
SEX
RED_CAR
REVOKED
CAR_USE
YOJ
TIF
0
2
4
6
%IncMSE
BLUEBOOK
AGE
HOME_VAL
INCOME
OLDCLAIM
TRAVTIME
CAR_TYPE
JOB
MVR_PTS
YOJ
TIF
CAR_AGE
EDUCATION
CLM_FREQ
HOMEKIDS
KIDSDRIV
URBANICITY
PARENT1
MSTATUS
CAR_USE
RED_CAR
REVOKED
SEX
0.0e+00
1.0e+10
IncNodePurity



PART 2: DATA PREPARATION

In the following section we will prepare and transform our variables for our model.

Clean Data

For incomplete cases, we replaced the value of `NULL` data with the mean of the relevant data vector.

We must also convert all of the education categories into numerical ones.

Bucket Transformations

Transform data by putting it into buckets.

Mathematical Transformations

Log or square root (or use Box-Cox) where applicable.

New Variables

Combine variables (such as ratios or adding or multiplying) to create new variables

PART 3. BUILD MODELS

Using the transformed data above, we developed two multiple linear regression and three binary logistic regression models. Through these models, we hope to predict **(1)** the probability that a person will crash their car and **(2)** the amount of money it will cost if the person does crash their car.

Multiple Linear Regression

MLR for `TARGET_AMT`.

MLR 1

[] Create Model #1
[] Describe the techniques you used. [] Show summarised results

MLR 2

[] Create Model #2
[] Describe the techniques you used. [] Show summarised results

MLR ANALYSIS

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?

Binary Logistic Regression

BLR for `TARGET_FLAG`

BLR 1

[] Create Model #1
[] Describe the techniques you used. [] Show summarised results

BLR 2

[] Create Model #2
[] Describe the techniques you used. [] Show summarised results

BLR 3

[] Create Model #3
[] Describe the techniques you used. [] Show summarised results

BLR ANALYSIS

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?

PART 4: SELECT MODELS

Select the best multiple linear regression model and the best binary logistic regression model. Discuss why you selected your models.

MLR Evaluation

Use a metric such as Adjusted R2, RMSE, etc. 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:

Mean Squared Error

R2

F-statistic

Residual plots.

BLR Evaluation

Use a metric such as log likelihood, AIC, ROC curve, etc. Using the training data set, evaluate the binary logistic regression model based on:

Accuracy

Classification Error Rate

Precision

Sensitivity

Specificity

F1 score

AUC

Confusion Matrix

Predictions

Make predictions using the evaluation data set.