HW #4

Critical Thinking Group One

2021-05-02

Contents

Authorship

Critical Thinking Group 1

- Angel Claudio
- Bonnie Cooper
- Manolis Manoli
- Magnus Skonberg
- Christian Thieme
- Leo Yi

Abstract

We will explore, analyze and model a data set containing approximately 8,000 records. Each row represents 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 will only use the variables given to us (or variables that we derive from the variables provided).

Data Exploration

Structure of Data

```
## Rows: 8,161
## Columns: 26
## $ INDEX       <dbl> 1, 2, 4, 5, 6, 7, 8, 11, 12, 13, 14, 15, 16, 17, 19, 20, 2~
## $ TARGET_FLAG <dbl> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <dbl> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <dbl> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <dbl> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", NA, "$114,986", "$125,301~
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No~
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "$0"~
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", "Yes",~
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", "M"~
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High School", "~
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Blue Colla~
## $ TRAVTIME    <dbl> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Private", ~
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000", "$17~
## $ TIF         <dbl> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV", "Sports~
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes", "no", ~
## $ OLDCLAIM    <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0", "$~
## $ CLM_FREQ    <dbl> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "N~
## $ MVR_PTS     <dbl> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <dbl> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly Urba~
```

Right away we can see we'll have some work to do. It appears that many features that are factors have imported as characters or doubles. It is also clear that there may be some ordinal levels within some of the factors. We'll note this for our data cleaning section. We also note that the training set has 26 columns and 8,161 rows.

Before transformations, lets see how many of our columns are numeric and how many are characters:

Numeric:

```
## [1] 12
```

Character:

```
## [1] 14
```

Data Metrics

Next, we'll quickly look at a summary of each of our features to quickly get a bird's eye view of our distributions:

```
##      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
##  Min.   :16.00   Min.   :0.0000   Min.   : 0.0   Length:8161
##  1st Qu.:39.00   1st Qu.:0.0000   1st Qu.: 9.0   Class :character
##  Median :45.00   Median :0.0000   Median :11.0   Mode  :character
##  Mean   :44.79   Mean   :0.7212   Mean   :10.5
##  3rd Qu.:51.00   3rd Qu.:1.0000   3rd Qu.:13.0
##  Max.   :81.00   Max.   :5.0000   Max.   :23.0
##  NA's   :6                        NA's   :454
##    PARENT1            HOME_VAL           MSTATUS              SEX
##  Length:8161        Length:8161        Length:8161        Length:8161
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##   EDUCATION             JOB              TRAVTIME        CAR_USE
##  Length:8161        Length:8161        Min.   :  5.00   Length:8161
##  Class :character   Class :character   1st Qu.: 22.00   Class :character
##  Mode  :character   Mode  :character   Median : 33.00   Mode  :character
##                                        Mean   : 33.49
##                                        3rd Qu.: 44.00
##                                        Max.   :142.00
##
##    BLUEBOOK              TIF           CAR_TYPE           RED_CAR
##  Length:8161        Min.   : 1.000   Length:8161        Length:8161
##  Class :character   1st Qu.: 1.000   Class :character   Class :character
##  Mode  :character   Median : 4.000   Mode  :character   Mode  :character
##                     Mean   : 5.351
##                     3rd Qu.: 7.000
##                     Max.   :25.000
##
##    OLDCLAIM           CLM_FREQ       REVOKED             MVR_PTS
##  Length:8161        Min.   :0.0000   Length:8161        Min.   : 0.000
##  Class :character   1st Qu.:0.0000   Class :character   1st Qu.: 0.000
##  Mode  :character   Median :0.0000   Mode  :character   Median : 1.000
##                     Mean   :0.7986                      Mean   : 1.696
##                     3rd Qu.:2.0000                      3rd Qu.: 3.000
##                     Max.   :5.0000                      Max.   :13.000
##
##     CAR_AGE        URBANICITY
##  Min.   :-3.000   Length:8161
##  1st Qu.: 1.000   Class :character
##  Median : 8.000   Mode  :character
##  Mean   : 8.328
##  3rd Qu.:12.000
```

```
## Max.   :28.000
## NA's   :510
```

Some notes of interest:

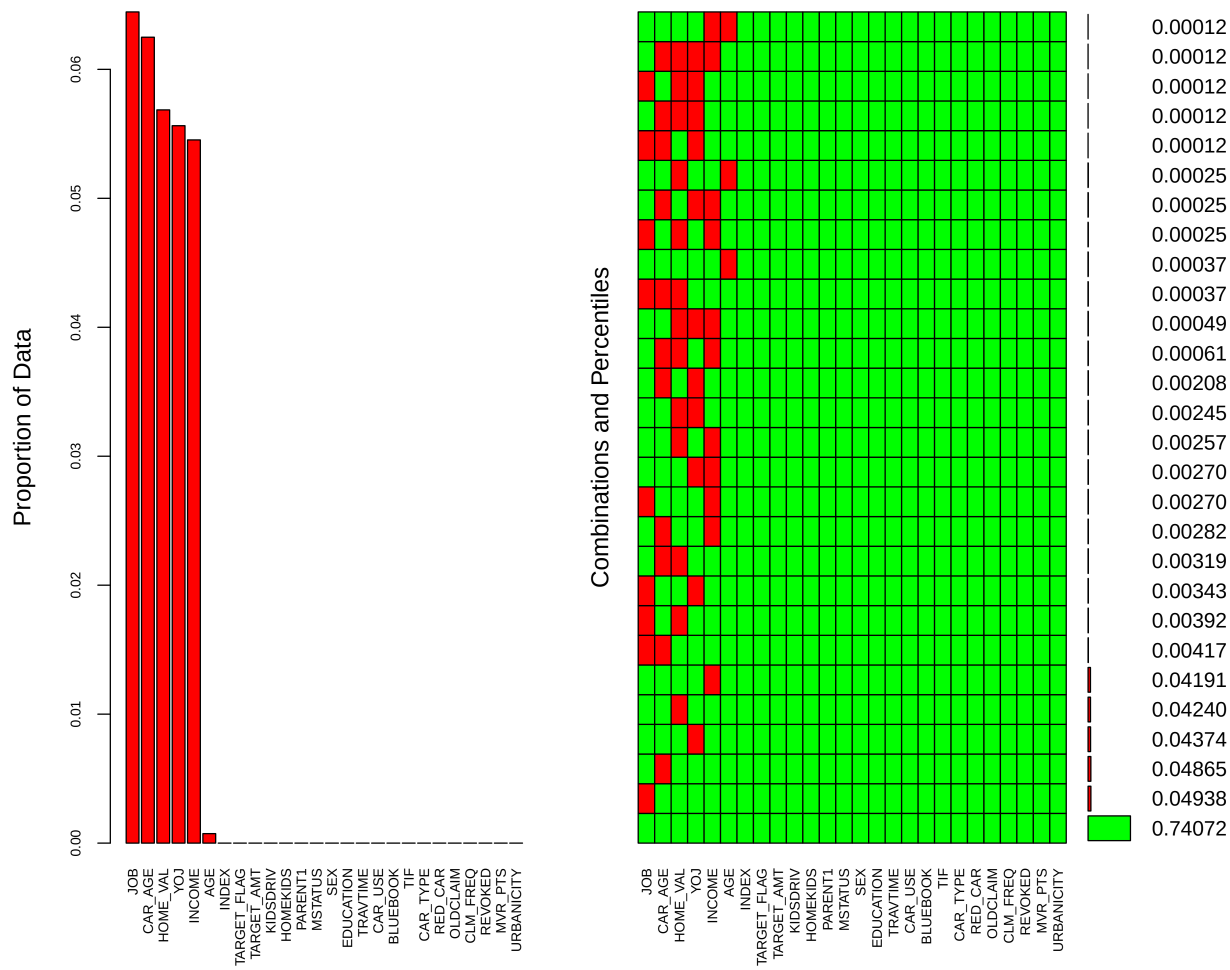
- KIDSDRIV: Max is 4
- AGE: minimum age is 16 which we'd expect and oldest individual is 81. There are 6 NA values
- HOMEKIDS: Max is 5
- TRAVTIME: It appears there may be some outliers here. 75% of the population is below 44 minutes. Max value is 142
- TIF: The majority of people are not long time customers
- CLM_FREQ: Maximum is over 5 years
- MVR_PTS: 75% have 3 or less, maximum is 13
- CAR_AGE: Appears to have some data that is wrong – shows minimum as -3. Max is 28

NA's Summary

Before going farther, let's see how extensive missing values are in our dataset.

```
##       INDEX TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS
##           0           0           0           0           6           0
##         YOJ      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX
##         454         445           0         464           0           0
##   EDUCATION         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF
##           0         526           0           0           0           0
##    CAR_TYPE     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS
##           0           0           0           0           0           0
##     CAR_AGE  URBANICITY
##         510           0
```

There are missing values in YOJ, INCOME, HOME_VAL, JOB, and CAR_AGE, however it does not appear to be pervasive. The feature with the most NA's, JOB, is only missing ~7% of its values. Let's see if the missing data is random in nature or if there is an underlying pattern.



We note that:

- Overall 74% of the data is free of missing values.
- Both JOB and CAR__AGE each represent almost 5% of NAs.
- As runner ups INCOME, HOME__VALU, and YOJ each represent about 4% of NAs.
- The low % of combinations seem to indicate that the NAs are random in nature.

We will address these NAs in the data transformation section.

Data Exploration Summary

The dataset has 27 variables and 8,161 observations. We summarize the following issues:

- NAs can be found in features like CAR__AGE and YOJ
- There appear to be outliers in several of the features
- There are character type data that should be numeric such as income
- There are character type data that should be factor such as marital status
- The index feature is not needed

Data Preparation

Exclusions

As part of our data preparation, we will drop the “Index” feature as there is no use for it.

Data After Exclusions

```
## Rows: 10,302
## Columns: 25
## $ TARGET_FLAG <dbl> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <dbl> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <dbl> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <dbl> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, 11, 0, 1~
## $ INCOME      <chr> "$67,349", "$91,449", "$16,039", NA, "$114,986", "$125,301~
## $ PARENT1     <chr> "No", "No", "No", "No", "No", "Yes", "No", "No", "No", "No~
## $ HOME_VAL    <chr> "$0", "$257,252", "$124,191", "$306,251", "$243,925", "$0"~
## $ MSTATUS     <chr> "z_No", "z_No", "Yes", "Yes", "Yes", "z_No", "Yes", "Yes",~
## $ SEX         <chr> "M", "M", "z_F", "M", "z_F", "z_F", "z_F", "M", "z_F", "M"~
## $ EDUCATION   <chr> "PhD", "z_High School", "z_High School", "<High School", "~
## $ JOB         <chr> "Professional", "z_Blue Collar", "Clerical", "z_Blue Colla~
## $ TRAVTIME    <dbl> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <chr> "Private", "Commercial", "Private", "Private", "Private", ~
## $ BLUEBOOK    <chr> "$14,230", "$14,940", "$4,010", "$15,440", "$18,000", "$17~
## $ TIF         <dbl> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <chr> "Minivan", "Minivan", "z_SUV", "Minivan", "z_SUV", "Sports~
## $ RED_CAR     <chr> "yes", "yes", "no", "yes", "no", "no", "no", "yes", "no", ~
## $ OLDCLAIM    <chr> "$4,461", "$0", "$38,690", "$0", "$19,217", "$0", "$0", "$~
## $ CLM_FREQ    <dbl> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <chr> "No", "No", "No", "No", "Yes", "No", "No", "Yes", "No", "N~
## $ MVR_PTS     <dbl> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <dbl> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <chr> "Highly Urban/ Urban", "Highly Urban/ Urban", "Highly Urba~
```

Character to Numeric Type

Additionally, we’ll change the following character type features to numeric:

- INCOME
- HOME_VAL
- OLDCLAIM
- BLUEBOOK

Features after Transformation from character to numeric

```
## Rows: 10,302
## Columns: 4
## $ INCOME   <dbl> 67349, 91449, 16039, NA, 114986, 125301, 18755, 107961, 62978~
## $ HOME_VAL <dbl> 0, 257252, 124191, 306251, 243925, 0, NA, 333680, 0, 0, 0, 20~
## $ BLUEBOOK <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 8780, 16970, 11200, ~
## $ OLDCLAIM <dbl> 4461, 0, 38690, 0, 19217, 0, 0, 2374, 0, 0, 0, 0, 5028, 0, 0,~
```

Character to Factor Type

We also identified that there would be numerous columns that would have to be transformed from a character data type to factor. Their final state is shown below after the transformation:

Features after Transformation from Character to Factor

```
## $MSTATUS
## [1] "No"  "Yes"
##
## $SEX
## [1] "F" "M"
##
## $JOB
## [1] "Blue Collar"  "Clerical"     "Doctor"       "Home Maker"   "Lawyer"
## [6] "Manager"      "Professional" "Student"
##
## $CAR_TYPE
## [1] "Minivan"     "Panel Truck" "Pickup"      "Sports Car"  "SUV"
## [6] "Van"
##
## $URBANICITY
## [1] "Highly Rural/ Rural" "Highly Urban/ Urban"
##
## $CAR_USE
## [1] "Commercial" "Private"
##
## $REVOKED
## [1] "No"  "Yes"
##
## $PARENT1
## [1] "No"  "Yes"
##
## $RED_CAR
## [1] "no"  "yes"
##
## $TARGET_FLAG
## [1] "0" "1"
```

Imputation

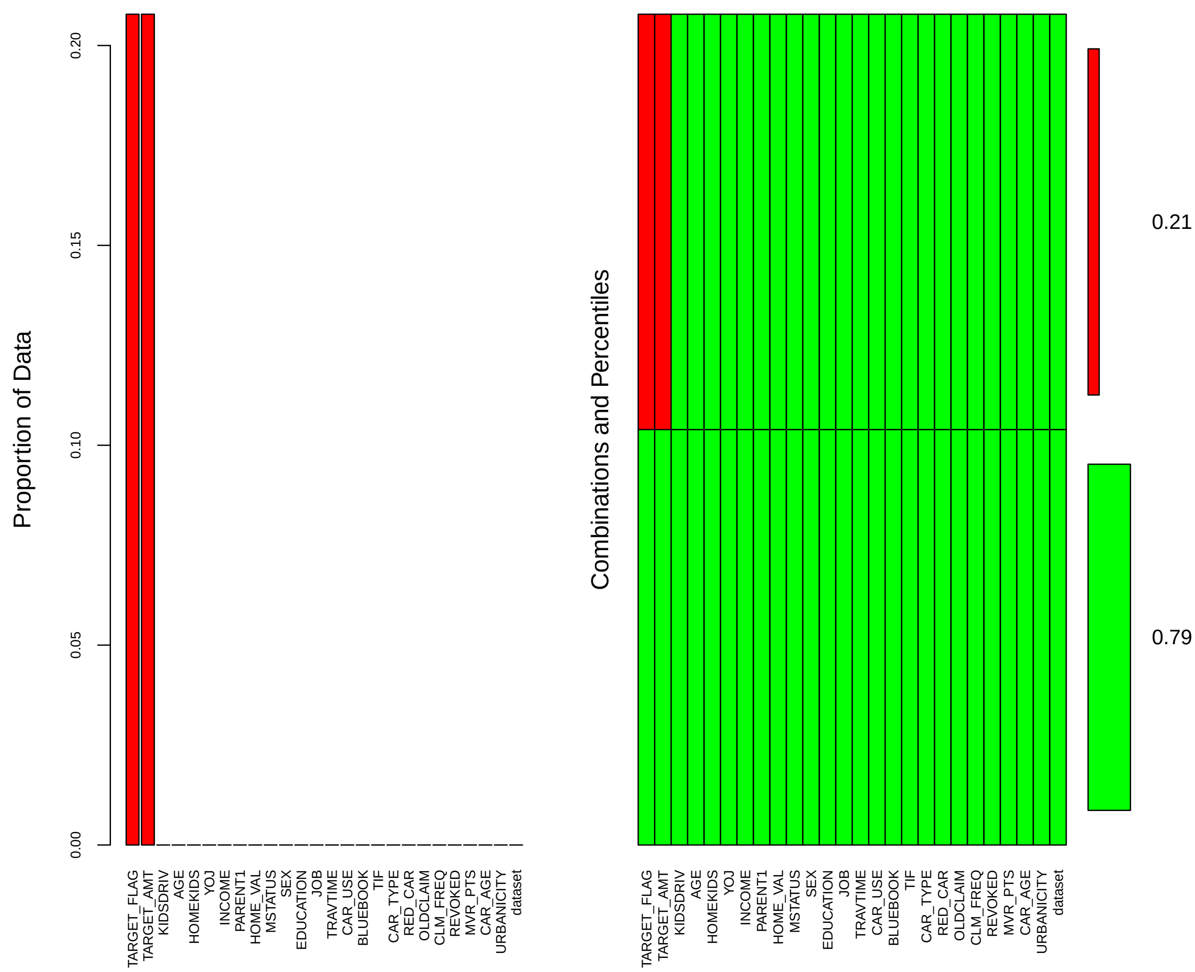
As noted in our EDA section, there are NAs in several of our features. Based on our analysis above, as there weren't any distinct patterns within missing data, we'll impute NAs using linear interpolation (`zoo` library) as well as filling in the blanks in JOBS with new level, 'Unknown'.

```
##       JOB
##  [1,] "Blue Collar"
##  [2,] "Clerical"
##  [3,] "Doctor"
##  [4,] "Home Maker"
##  [5,] "Lawyer"
##  [6,] "Manager"
##  [7,] "Professional"
```

```
##  [8,] "Student"
##  [9,] "Unknown"
```

Results Post-Imputation

```
## TARGET_FLAG  TARGET_AMT    KIDSDRIV         AGE    HOMEKIDS         YOJ
##        2141        2141           0           0           0           0
##      INCOME     PARENT1    HOME_VAL     MSTATUS         SEX   EDUCATION
##           0           0           0           0           0           0
##         JOB    TRAVTIME     CAR_USE    BLUEBOOK         TIF    CAR_TYPE
##           0           0           0           0           0           0
##     RED_CAR    OLDCLAIM    CLM_FREQ     REVOKED     MVR_PTS     CAR_AGE
##           0           0           0           0           0           0
##  URBANICITY     dataset
##           0           0
```



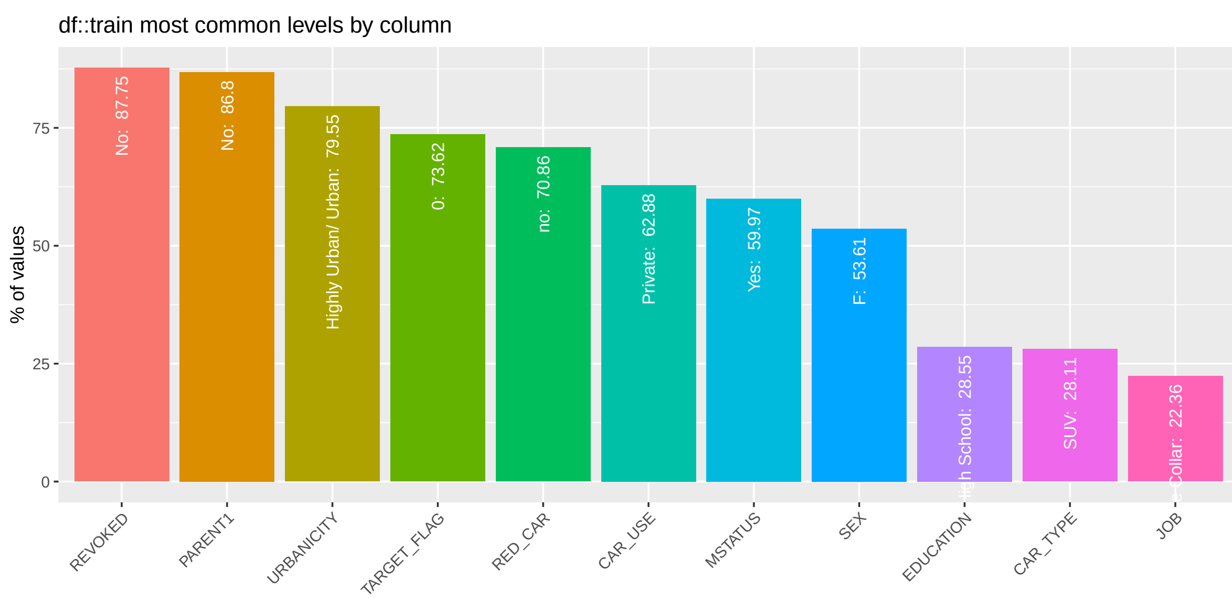
Data After Clean Up

```
## Rows: 10,302
## Columns: 25
```

```
## $ TARGET_FLAG <fct> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0, 0, 0, 1~
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000, 4021.0~
## $ KIDSDRIV    <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ AGE         <dbl> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55, 53, 45~
## $ HOMEKIDS    <dbl> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1~
## $ YOJ         <dbl> 11.00000, 11.00000, 10.00000, 14.00000, 13.00000, 12.00000~
## $ INCOME      <dbl> 67349.0, 91449.0, 16039.0, 65512.5, 114986.0, 125301.0, 18~
## $ PARENT1     <fct> No, No, No, No, No, Yes, No, No, No, No, No, No, No, No, N~
## $ HOME_VAL    <dbl> 0, 257252, 124191, 306251, 243925, 0, 166840, 333680, 0, 0~
## $ MSTATUS     <fct> No, No, Yes, Yes, Yes, No, Yes, Yes, No, No, No, Yes, Yes,~
## $ SEX         <fct> M, M, F, M, F, F, F, M, F, M, F, F, M, M, F, F, M, F, F, F~
## $ EDUCATION   <fct> PhD, High School, High School, <High School, PhD, Bachelor~
## $ JOB         <fct> Professional, Blue Collar, Clerical, Blue Collar, Doctor, ~
## $ TRAVTIME    <dbl> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25, 64, 48,~
## $ CAR_USE     <fct> Private, Commercial, Private, Private, Private, Commercial~
## $ BLUEBOOK    <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 8780, 16970, 1120~
## $ TIF         <dbl> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 6, 7, 4, ~
## $ CAR_TYPE    <fct> Minivan, Minivan, SUV, Minivan, SUV, Sports Car, SUV, Van,~
## $ RED_CAR     <fct> yes, yes, no, yes, no, no, no, yes, no, no, no, no, yes, y~
## $ OLDCLAIM    <dbl> 4461, 0, 38690, 0, 19217, 0, 0, 2374, 0, 0, 0, 0, 5028, 0,~
## $ CLM_FREQ    <dbl> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 0, 2~
## $ REVOKED     <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, No, Yes, No,~
## $ MVR_PTS     <dbl> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0, 0, 0, ~
## $ CAR_AGE     <dbl> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5, 13, 16,~
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly Urban/ Ur~
```

Factor Analysis

Let's take a look at the makeup of our factor variables. We'll look at the most common category within these variables.

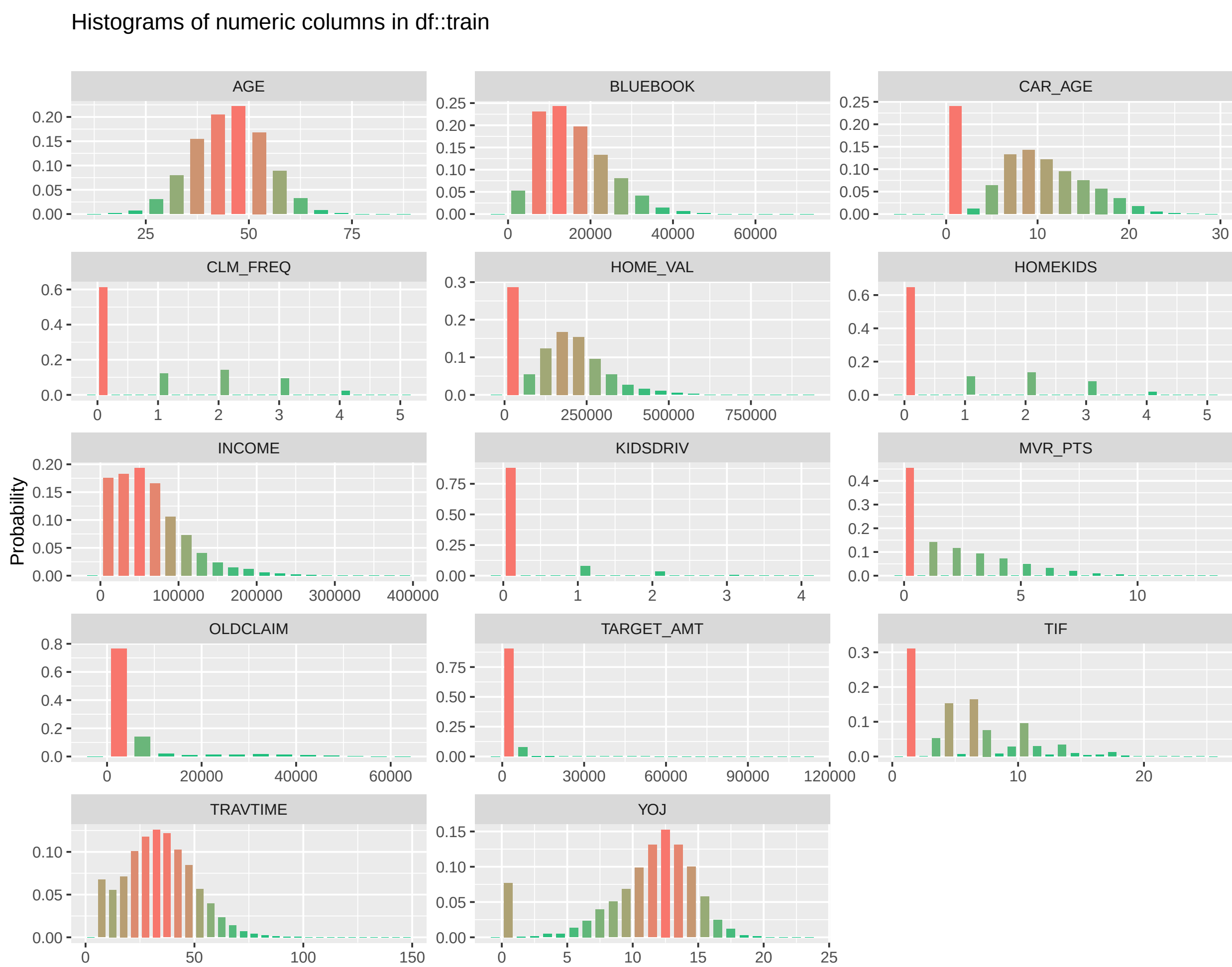


In looking at the above output, notes of interest are:

- 12% of the population has had their license revoked within the last 7 years (% seems high)

- A majority of the data comes from people in highly urban or urban areas
- 27% of the population has been involved in an accident
- Most cars aren't red
- Majority of the data is from private care use
- More than half of the population is married
- Almost a 50/50 split on M vs F
- SUV's make up more than 1/4 of the dataset
- Blue collar workers make up ~1/4 of the data

Summary of Common Levels



In looking at the plots above, we note:

- AGE looks normally distributed, which we'd expect
- BLUEBOOK values are right skewed. This may mean there is a correlation between the income of the individual and the type of education and job they have
- CAR_AGE: This one seems off. It looks like 25% of cars are new, yet the BLUEBOOK values are pretty low. Seems it would be difficult to buy a "new" car at such low values
- EDUCATION: 35%+ have less than a bachelor's degree
- HOME_VAL: It appears home value of 0 is pretty frequent. Instead of having a home with $0 value, this probably indicates they don't own a home. We'll need to clean this up and create a categorical

variable to capture this
- HOMEKIDS: Most people don't have kids
- INCOME: Income is very right skewed. Looks like over 60% of the population makes $50K or less
- KIDSDRIV: Majority of people don't have kids at home that are driving
- TIF: Appears to be a multi-modal distribution, possibly indicating sub-populations or errors within the data. 30%+ are new customers
- TRAVTIME: This is a bi-modal distribution. We may need to explore if there is a sub-population here.
- YOJ: Another bi-modal distribution. Are these distributions related to age and people becoming old enough to drive (turning 16 years old)?
- TARGET_AMT: Heavily right skewed. We'll look in our transformation section to see if this target variable could benefit from a transformation

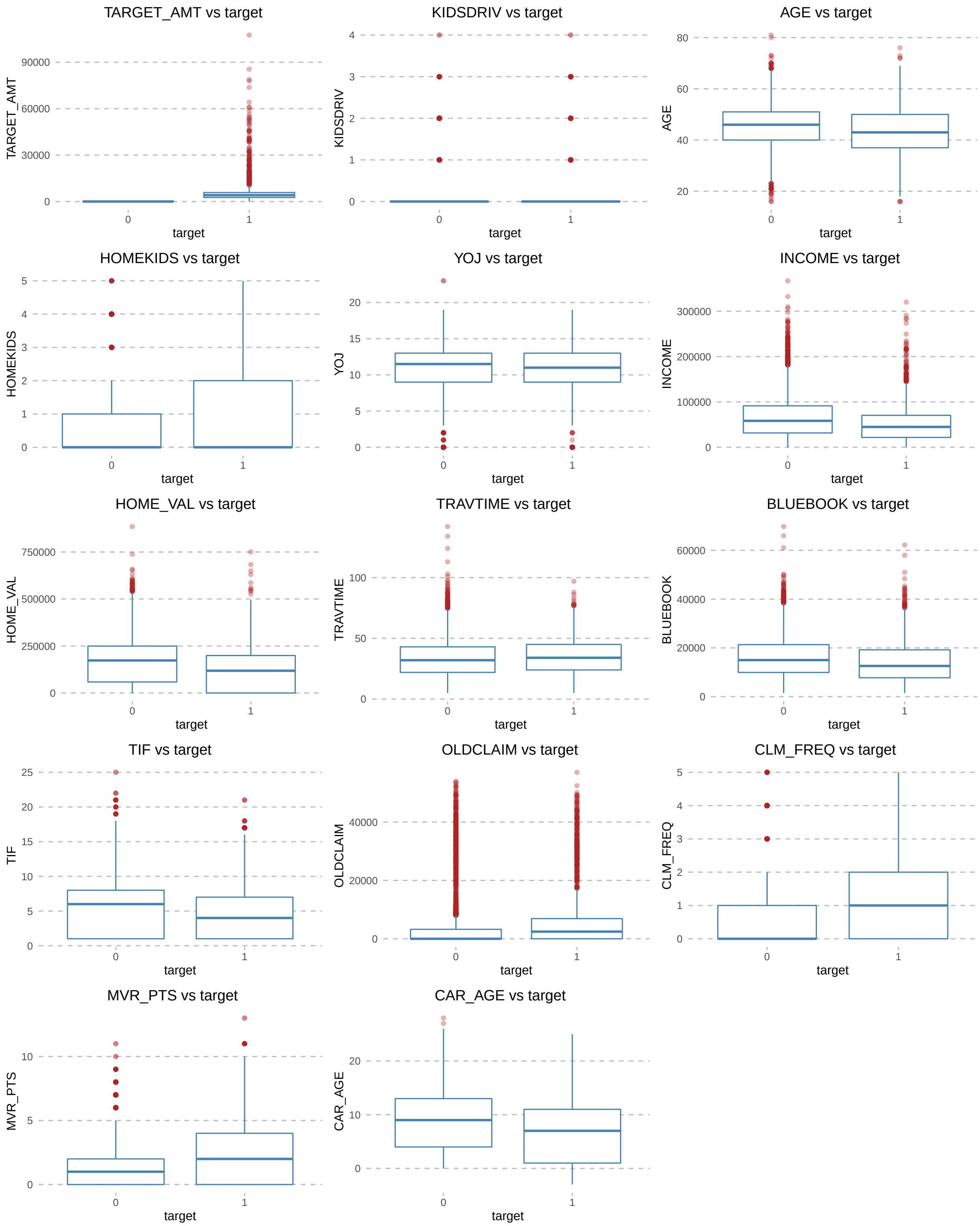
We can see above that many of these features have 0s indicating that the variable does not have a value for the field. We'll need to make some categorical features to capture this signal in our feature engineering section.

Now let's look at the relationship between our variables and TARGET_FLAG

Boxplots for when Predictor is TARGET_FLAG

```
# target_name <- 'your_target_name'
target_name <- 'TARGET_FLAG'

boxplot_depend_vs_independ(train, target_name)
```



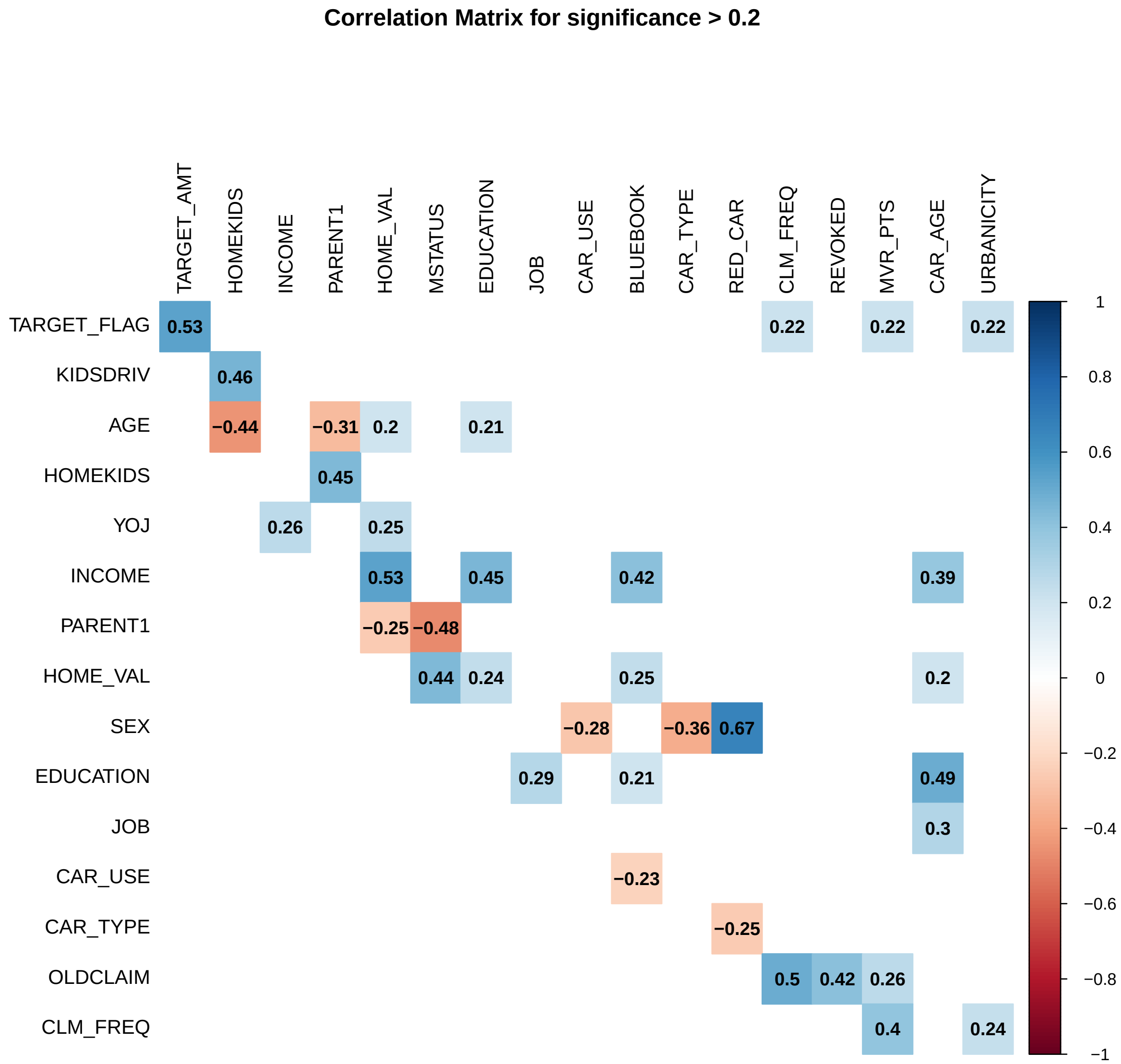
In looking at the above boxplots, we note the following:

- KIDSDRIVE appears to have no relationship with TARGET_FLAG
- AGE appears to have a weak relationship with TARTGET_FLAG. Those with a lower age, on average, get in more wrecks

- HOMEKIDS doesn't appear to have a meaningful relationship
- YOJ does look like it has a weak relationship with TARGET_FLAG, the less YOJ, the more wrecks
- INCOME also appears to have a relationship, the lower income, the more likely to wreck
- HOME_VAL appears to have a relationship and is probably somewhat skewed because of the 0s in the distribution
- TRAVTIME looks to have a weak relationship, with those who have longer travel times being more likely to wreck
- BLUEBOOK has a relationship as well, with those with lower bluebook values being more likely to wreck
- TIF appears to have a relationship with those who have been customers longer being less likely to wreck
- OLDCLAIM appears to have a relationship and those who have higher claim values are more likely to wreck
- CLAIM_FREQ definitely has a relationship
- MVR_PTS has a strong relationship as well with those with more points being more likely to wreck
- CAR_AGE seems to have a strong relationship as well with older car owners being less likely to wreck

Having looked at our features, let's now take a look at a correlation plot to see the strength between our variables.

Correlation Matrix



As noted previously, multi-collinearity is not a huge issue with this dataset. We note the following variables that are collinear:

- HOMEKIDS and KIDSDRIV
- AGE and HOMEKIDS
- INCOME and HOME_VAL
- INCOME and BLUEBOOK and CAR_AGE
- CLM_FREQ and OLDCLAIM
- CLM_FREQ and MVR_PTS

Feature Engineering

From observations made from the data exploration and preparation above, we create several flag variables to captured observed trends in the data:

1. Brand New Car Flag
2. Zero Claims History Flag
3. Home Ownership Flag
4. Clean Motor Vehicle Record Flag
5. Years on Job Flag

```
## Rows: 10,302
## Columns: 31
## $ TARGET_FLAG           <fct> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0,~
## $ TARGET_AMT            <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.~
## $ KIDSDRIV              <dbl> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ AGE                   <dbl> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43,~
## $ HOMEKIDS              <dbl> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3,~
## $ YOJ                   <dbl> 11.00000, 11.00000, 10.00000, 14.00000, 13.0000~
## $ INCOME                <dbl> 67349.0, 91449.0, 16039.0, 65512.5, 114986.0, 1~
## $ PARENT1               <fct> No, No, No, No, No, Yes, No, No, No, No, No, No~
## $ HOME_VAL              <dbl> 0, 257252, 124191, 306251, 243925, 0, 166840, 3~
## $ MSTATUS               <fct> No, No, Yes, Yes, Yes, No, Yes, Yes, No, No, No~
## $ SEX                   <fct> M, M, F, M, F, F, F, M, F, M, F, F, M, M, F, F,~
## $ EDUCATION             <fct> PhD, High School, High School, <High School, Ph~
## $ JOB                   <fct> Professional, Blue Collar, Clerical, Blue Colla~
## $ TRAVTIME              <dbl> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, ~
## $ CAR_USE               <fct> Private, Commercial, Private, Private, Private,~
## $ BLUEBOOK              <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 8780, ~
## $ TIF                   <dbl> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6~
## $ CAR_TYPE              <fct> Minivan, Minivan, SUV, Minivan, SUV, Sports Car~
## $ RED_CAR               <fct> yes, yes, no, yes, no, no, no, yes, no, no, no,~
## $ OLDCLAIM              <dbl> 4461, 0, 38690, 0, 19217, 0, 0, 2374, 0, 0, 0, ~
## $ CLM_FREQ              <dbl> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0,~
## $ REVOKED               <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, N~
## $ MVR_PTS               <dbl> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0~
## $ CAR_AGE               <dbl> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10,~
## $ URBANICITY            <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highl~
## $ dataset               <chr> "train", "train", "train", "train", "train", "t~
## $ CAR_AGE_BRAND_NEW_FLAG <dbl> 0, 1, 0, 0, 0, 0, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0,~
## $ CLM_FREQ_ZERO         <dbl> 0, 1, 0, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 1, 1, 1,~
## $ HOME_VAL_ZERO         <dbl> 1, 0, 0, 0, 0, 1, 0, 0, 1, 1, 1, 0, 0, 1, 0, 0,~
## $ MVR_PTS_ZERO          <dbl> 0, 1, 0, 1, 0, 1, 1, 0, 1, 0, 1, 1, 0, 0, 0, 1,~
## $ YOJ_ZERO              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0,~
```

Build Models

From the insights we gained from EDA, we move forward by implementing two main modeling approaches. We begin by splitting the training data into train and validation sets. For our first modeling approach, we use `TARGET_FLAG` as a binary response variable in conjunction with the original and engineered featured variables of our processed data. The second modeling approach uses the `TARGET_AMT` response variable. `TARGET_AMT` is a continuous numeric feature which we use to deploy multiple linear regression models. We developed several models for both the binary regression and multivariate linear regression approaches and evaluate each model's performance to select the best model.

Modeling the Binary Response Variable

Here we describe binomial modeling that utilizes the feature set to predict the binary response variable, `TARGET_FLAG`. Where `TARGET_FLAG` coded '1' is a car that was in a crash and '0' otherwise.

Model #1: Binary Logistic Model

To find a baseline for performance with the Binary Response Variable, we begin with a binary logistic regression model that includes all feature variables (original and engineered).

Model 1 Summary:

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = partial_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6112  -0.7066  -0.3873   0.6071   3.1642
##
## Coefficients:
##                                   Estimate    Std. Error z value Pr(>|z|)
## (Intercept)                    -2.6353852570  0.3692938248  -7.136 9.59e-13 ***
## KIDSDRIV                        0.3571878954  0.0672347452   5.313 1.08e-07 ***
## AGE                            -0.0044821229  0.0044437592  -1.009 0.313151
## HOMEKIDS                        0.0283779232  0.0413952129   0.686 0.493006
## YOJ                             0.0199859786  0.0131421674   1.521 0.128322
## INCOME                         -0.0000038983  0.0000012297  -3.170 0.001524 **
## PARENT1Yes                      0.4345261543  0.1195415167   3.635 0.000278 ***
## HOME_VAL                       -0.0000008135  0.0000006038  -1.347 0.177869
## MSTATUSYes                     -0.5570080425  0.0945118876  -5.894 3.78e-09 ***
## SEXM                            0.0574919448  0.1233997371   0.466 0.641287
## EDUCATIONBachelors             -0.3237731474  0.1259104667  -2.571 0.010127 *
## EDUCATIONHigh School            0.0477183952  0.1039762676   0.459 0.646281
## EDUCATIONMasters               -0.2228806421  0.1949540289  -1.143 0.252936
## EDUCATIONPhD                   -0.1578312765  0.2344554502  -0.673 0.500831
## JOBClerical                     0.1526848730  0.1156030866   1.321 0.186579
## JOBDoctor                      -0.8927033480  0.3147629808  -2.836 0.004567 **
## JOBHome Maker                  -0.2374443280  0.1748996401  -1.358 0.174590
## JOBLawyer                      -0.2610547204  0.2031345102  -1.285 0.198746
## JOBManager                     -0.8825723265  0.1505595628  -5.862 4.57e-09 ***
## JOBProfessional                -0.1821565905  0.1305048683  -1.396 0.162780
## JOBStudent                     -0.3533123133  0.1548680858  -2.281 0.022526 *
## JOBUnknown                     -0.3575077197  0.2012075210  -1.777 0.075599 .
## TRAVTIME                        0.0163233345  0.0020739450   7.871 3.53e-15 ***
## CAR_USEPrivate                 -0.7573111596  0.1001512340  -7.562 3.98e-14 ***
## BLUEBOOK                       -0.0000203253  0.0000057075  -3.561 0.000369 ***
## TIF                            -0.0602765321  0.0080225726  -7.513 5.76e-14 ***
## CAR_TYPEPanel Truck             0.6249326429  0.1754299365   3.562 0.000368 ***
## CAR_TYPEPickup                  0.5892766952  0.1114515847   5.287 1.24e-07 ***
```

```
## CAR_TYPESports Car             1.1304010611  0.1423940203   7.939 2.05e-15 ***
## CAR_TYPESUV                    0.8307123573  0.1223938619   6.787 1.14e-11 ***
## CAR_TYPEVan                    0.5957753869  0.1389737506   4.287 1.81e-05 ***
## RED_CARyes                     0.0739752125  0.0949719070   0.779 0.436029
## OLDCLAIM                      -0.0000204619  0.0000045945  -4.454 8.44e-06 ***
## CLM_FREQ                       0.0263147937  0.0484439230   0.543 0.586991
## REVOKEDYes                     0.9095662092  0.1017607186   8.938  < 2e-16 ***
## MVR_PTS                        0.1046492368  0.0209733902   4.990 6.05e-07 ***
## CAR_AGE                        0.0136207721  0.0112026232   1.216 0.224040
## URBANICITYHighly Urban/ Urban  2.4130546748  0.1229448424  19.627  < 2e-16 ***
## CAR_AGE_BRAND_NEW_FLAG         0.1738136434  0.1143758529   1.520 0.128594
## CLM_FREQ_ZERO                 -0.5642737003  0.1324463830  -4.260 2.04e-05 ***
## HOME_VAL_ZERO                  0.1108424845  0.1510345261   0.734 0.463017
## MVR_PTS_ZERO                   0.0549228747  0.0932620539   0.589 0.555922
## YOJ_ZERO                       0.7488766866  0.2114227249   3.542 0.000397 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8007.9  on 6937  degrees of freedom
## Residual deviance: 6130.9  on 6895  degrees of freedom
## AIC: 6216.9
##
## Number of Fisher Scoring iterations: 5
```

We can see the AIC result from a binomial model using the logit link function. To invoke a parsimonious approach, another model will be derived using a stepwise method to further narrow down models based on significance and possible negative effects of multicollinearity.

Model #2: Stepwise Binary Logistic Model

Using the full model as a starting point, the following summarizes the output of a backward stepwise by AIC model selection process:

```
##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + REVOKED + MVR_PTS +
##     URBANICITY + CLM_FREQ_ZERO + YOJ_ZERO, family = binomial,
##     data = partial_train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5898  -0.7121  -0.3868   0.6059   3.1691
##
## Coefficients:
##                                     Estimate    Std. Error z value Pr(>|z|)
## (Intercept)                   -2.4244763223  0.2662714833  -9.105  < 2e-16 ***
## KIDSDRIV                       0.3732736585  0.0602428662   6.196 5.79e-10 ***
## YOJ                            0.0205002704  0.0125316736   1.636 0.101865
## INCOME                        -0.0000034578  0.0000011281  -3.065 0.002176 **
```

```
## PARENT1Yes                      0.5077903370  0.1024822079   4.955 7.24e-07 ***
## HOME_VAL                       -0.0000012228  0.0000003604  -3.393 0.000692 ***
## MSTATUSYes                     -0.5519465230  0.0878475807  -6.283 3.32e-10 ***
## EDUCATIONBachelors             -0.3193276884  0.1190011519  -2.683 0.007288 **
## EDUCATIONHigh School            0.0551862180  0.1032493480   0.534 0.592999
## EDUCATIONMasters               -0.1678285153  0.1749151756  -0.959 0.337314
## EDUCATIONPhD                   -0.1183186760  0.2184556072  -0.542 0.588084
## JOBClerical                     0.1574119985  0.1150454747   1.368 0.171231
## JOBDoctor                      -0.8894686944  0.3140531691  -2.832 0.004623 **
## JOBHome Maker                  -0.2796037981  0.1727264094  -1.619 0.105497
## JOBLawyer                      -0.2793048471  0.2027981223  -1.377 0.168433
## JOBManager                     -0.8978030846  0.1501909597  -5.978 2.26e-09 ***
## JOBProfessional                -0.1910254808  0.1300858481  -1.468 0.141980
## JOBStudent                     -0.3113797704  0.1476631446  -2.109 0.034969 *
## JOBUnknown                     -0.3636478361  0.2009987786  -1.809 0.070419 .
## TRAVTIME                        0.0162119532  0.0020684039   7.838 4.58e-15 ***
## CAR_USEPrivate                 -0.7544150738  0.0999990582  -7.544 4.55e-14 ***
## BLUEBOOK                       -0.0000233274  0.0000051233  -4.553 5.28e-06 ***
## TIF                            -0.0595111446  0.0079996107  -7.439 1.01e-13 ***
## CAR_TYPEPanel Truck             0.6900331622  0.1634189761   4.222 2.42e-05 ***
## CAR_TYPEPickup                  0.5863078702  0.1111672085   5.274 1.33e-07 ***
## CAR_TYPESports Car              1.0522835531  0.1171887863   8.979  < 2e-16 ***
## CAR_TYPESUV                     0.7547530556  0.0945097045   7.986 1.39e-15 ***
## CAR_TYPEVan                     0.6323387541  0.1341998064   4.712 2.45e-06 ***
## OLDCLAIM                       -0.0000204398  0.0000045839  -4.459 8.23e-06 ***
## REVOKEDYes                      0.9091169573  0.1015518485   8.952  < 2e-16 ***
## MVR_PTS                         0.0970773926  0.0153772677   6.313 2.74e-10 ***
## URBANICITYHighly Urban/ Urban  2.4159661906  0.1229882675  19.644  < 2e-16 ***
## CLM_FREQ_ZERO                  -0.6188092494  0.0854268151  -7.244 4.36e-13 ***
## YOJ_ZERO                        0.7665186259  0.2061649545   3.718 0.000201 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8007.9  on 6937  degrees of freedom
## Residual deviance: 6138.1  on 6904  degrees of freedom
## AIC: 6206.1
##
## Number of Fisher Scoring iterations: 5
```

The stepwise model by AIC reduces the dimensionality of the model and results in a more parsimonious fit of the data. For example, Model 1 involves 28 feature variables and fits 43 coefficients whereas the stepwise Model 2 uses 19 feature variables and fits 34 model coefficients. Additionally, we see Model 2 AIC is lower indicating lower estimated prediction error. This suggests that, in addition to being a simple model, the stepwise method works better to create an overall better fit to the data.

In looking at the coefficients, most of these make sense intuitively. What is interesting is how big a difference URBANICITY makes to the log odd percent. It does makes sense that individuals who live in highly urban areas would be involved in more accidents since there are so many more people on the roads.

Model 3: Random Forest Model

There are other approaches to modeling a binary response variable other than using binary linear regression. Here, we describe the fit of a random forest model that levies the same full feature variable set as Model 1

```
##
## Call:
##  randomForest(formula = TARGET_FLAG ~ ., data = partial_train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 5
##
##         OOB estimate of  error rate: 21.46%
## Confusion matrix:
##      0   1 class.error
## 0 4772 335  0.06559624
## 1 1154 677  0.63025669
```

Multivariate Regression Model

Here we describe several multivariate regression modeling efforts to utilize the feature set to predict the continuous numeric response variable, `TARGET_AMT`. `TARGET_AMT` gives the monetary amount of costs incurred if a car was involved in a crash.

Model 4 - Multiple Linear Regression

To begin multivariate methods, we model the response variable using the full set of feature variables

```
##
## Call:
## lm(formula = .outcome ~ ., data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -9551  -3185  -1528    478  98590
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     5754.27     184.12  31.252  < 2e-16 ***
## KIDSDRIV                         -67.17     222.46  -0.302  0.76274
## AGE                              386.33     230.67   1.675  0.09415 .
## HOMEKIDS                         484.47     281.96   1.718  0.08592 .
## YOJ                             -213.64     351.35  -0.608  0.54323
## INCOME                          -158.85     336.09  -0.473  0.63653
## PARENT1Yes                       -38.49     271.55  -0.142  0.88731
## HOME_VAL                        -171.84     426.98  -0.402  0.68739
## MSTATUSYes                      -566.42     281.82  -2.010  0.04460 *
## SEXM                             718.12     361.55   1.986  0.04716 *
## EDUCATIONBachelors                54.12     301.45   0.180  0.85754
## `EDUCATIONHigh School`          -113.91     275.04  -0.414  0.67880
## EDUCATIONMasters                 595.82     434.53   1.371  0.17049
## EDUCATIONPhD                     669.43     343.02   1.952  0.05115 .
```

```
## JOBClerical                          -10.65     245.64  -0.043  0.96544
## JOBDoctor                           -284.33     243.11  -1.170  0.24232
## `JOBHome Maker`                     -127.94     285.18  -0.449  0.65375
## JOBLawyer                             47.05     338.24   0.139  0.88939
## JOBManager                          -260.24     246.52  -1.056  0.29127
## JOBProfessional                      350.05     242.81   1.442  0.14958
## JOBStudent                            53.96     284.14   0.190  0.84941
## JOBUnknown                            43.09     312.10   0.138  0.89021
## TRAVTIME                              37.01     187.69   0.197  0.84369
## CAR_USEPrivate                      -313.76     290.04  -1.082  0.27950
## BLUEBOOK                             895.95     282.84   3.168  0.00156 **
## TIF                                  -61.87     186.30  -0.332  0.73987
## `CAR_TYPEPanel Truck`               -336.28     295.68  -1.137  0.25557
## CAR_TYPEPickup                       -51.94     269.15  -0.193  0.84699
## `CAR_TYPESports Car`                 375.79     294.36   1.277  0.20189
## CAR_TYPESUV                          323.29     348.25   0.928  0.35337
## CAR_TYPEVan                           50.84     250.89   0.203  0.83944
## RED_CARyes                          -223.90     250.36  -0.894  0.37129
## OLDCLAIM                             218.85     276.64   0.791  0.42899
## CLM_FREQ                            -317.51     330.98  -0.959  0.33753
## REVOKEDYes                          -560.44     238.66  -2.348  0.01897 *
## MVR_PTS                              157.17     266.09   0.591  0.55482
## CAR_AGE                            -1167.84     377.26  -3.096  0.00199 **
## `URBANICITYHighly Urban/ Urban`      154.71     189.07   0.818  0.41331
## CAR_AGE_BRAND_NEW_FLAG              -718.95     306.33  -2.347  0.01904 *
## CLM_FREQ_ZERO                        -73.59     371.22  -0.198  0.84288
## HOME_VAL_ZERO                       -505.51     419.18  -1.206  0.22799
## MVR_PTS_ZERO                         -23.69     253.30  -0.094  0.92551
## YOJ_ZERO                            -362.54     385.99  -0.939  0.34773
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7883 on 1790 degrees of freedom
## Multiple R-squared:  0.03589,    Adjusted R-squared:  0.01326
## F-statistic: 1.586 on 42 and 1790 DF,  p-value: 0.01014
```

As we can see from the summary output, this approach does not yield a statistically significant fit to the data and has a very low R^2 value. Model 4 gives a very poor fit to the data, but perhaps this can be improved upon.

MV Model 5 - Stepwise Multiple Linear Regression

Using Model 4 as a starting point, here we describe the results of a backwards stepwise by AIC multiple linear regression model selection process.

```
##
## Call:
## lm(formula = TARGET_AMT ~ AGE + HOMEKIDS + MSTATUS + SEX + BLUEBOOK +
##     REVOKED + CAR_AGE + CAR_AGE_BRAND_NEW_FLAG + HOME_VAL_ZERO,
##     data = partial_train_mv)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
##  -9219  -3180  -1568    436 100185
##
## Coefficients:
##                         Estimate Std. Error t value  Pr(>|t|)
## (Intercept)            4106.0624  1185.9980   3.462  0.000548 ***
## AGE                      44.3494    21.7557   2.039  0.041643 *
## HOMEKIDS                282.7888   172.2978   1.641  0.100912
## MSTATUSYes            -1096.9312   450.5195  -2.435  0.014995 *
## SEXM                    678.7742   374.2296   1.814  0.069874 .
## BLUEBOOK                  0.1000     0.0229   4.367 0.0000133 ***
## REVOKEDYes             -993.1943   459.4666  -2.162  0.030777 *
## CAR_AGE                -120.0143    53.8415  -2.229  0.025933 *
## CAR_AGE_BRAND_NEW_FLAG -1124.6087   626.0132  -1.796  0.072587 .
## HOME_VAL_ZERO          -756.9170   459.9343  -1.646  0.099997 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7863 on 1823 degrees of freedom
## Multiple R-squared:  0.0231, Adjusted R-squared:  0.01828
## F-statistic:  4.79 on 9 and 1823 DF,  p-value: 0.000002459
```

The stepwise model selection resulted in a statistically significant p-value. However, the R^2 value indicates the Model 5 does not describe much variability in the data. So far, the quality of the multivariate linear regression approaches have left much to be desired. Next, we perform transformations to select feature variables in an effort to improve the fit.

Model 6 - Multivariate Linear Regression with Box-Cox Transformations

Here we describe a model built with the intention of testing box cox transformations on non-normally distributed variables, with model selection based on a manual selection process, including only significant independent feature variables.

```
## Estimated transformation parameters
##         AGE    BLUEBOOK     CAR_AGE    CLM_FREQ    HOME_VAL    HOMEKIDS
## -3.26296061  0.45557345  0.53194959  0.19001232  0.95680511  0.74940129
##      INCOME    KIDSDRIV     MVR_PTS    OLDCLAIM  TARGET_AMT         TIF
##  0.88488262 -1.57276438  0.20739359 -0.17762378  0.01363055 -0.02223101
##    TRAVTIME         YOJ
##  0.76666788  1.72079913
```

```
##
## Call:
## lm(formula = log(TARGET_AMT) ~ . + I(BLUEBOOK^0.5) + I(MVR_PTS^0.33) +
##     I(CAR_AGE^0.5) + I(CLM_FREQ^0.33), data = partial_train_mv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0825 -0.4017  0.0406  0.4006  3.2382
##
## Coefficients:
##                                 Estimate    Std. Error t value Pr(>|t|)
## (Intercept)                 7.3503267097  0.9481804576   7.752 1.51e-14 ***
## KIDSDRIV                   -0.0229075710  0.0363465752  -0.630 0.528609
```

```
## AGE                            0.0034095676  0.0024432740   1.395 0.163041
## HOMEKIDS                       0.0388800594  0.0241293874   1.611 0.107288
## YOJ                           -0.0143172874  0.0079776891  -1.795 0.072876 .
## INCOME                        -0.0000004264  0.0000008095  -0.527 0.598400
## PARENT1Yes                    -0.0080604512  0.0671577670  -0.120 0.904479
## HOME_VAL                      -0.0000002183  0.0000003696  -0.591 0.554883
## MSTATUSYes                    -0.1173962098  0.0572854150  -2.049 0.040577 *
## SEXM                           0.0563239618  0.0744274068   0.757 0.449291
## EDUCATIONBachelors            -0.0778961132  0.0725655348  -1.073 0.283210
## EDUCATIONHigh School           0.0013253906  0.0584803896   0.023 0.981921
## EDUCATIONMasters               0.2108908131  0.1237628596   1.704 0.088557 .
## EDUCATIONPhD                   0.3209255664  0.1488179965   2.156 0.031178 *
## JOBClerical                    0.0036365167  0.0666575976   0.055 0.956499
## JOBDoctor                     -0.0479935527  0.2091021794  -0.230 0.818489
## JOBHome Maker                 -0.0741275869  0.1035850942  -0.716 0.474319
## JOBLawyer                     -0.0670703273  0.1338682125  -0.501 0.616420
## JOBManager                    -0.0408282169  0.1041269248  -0.392 0.695031
## JOBProfessional                0.0694767589  0.0771279076   0.901 0.367817
## JOBStudent                     0.0680772116  0.0876298039   0.777 0.437337
## JOBUnknown                    -0.0121676691  0.1307995491  -0.093 0.925894
## TRAVTIME                      -0.0008597352  0.0012380492  -0.694 0.487505
## CAR_USEPrivate                -0.0076256346  0.0590807243  -0.129 0.897316
## BLUEBOOK                      -0.0000429770  0.0000142395  -3.018 0.002579 **
## TIF                           -0.0023724797  0.0048457051  -0.490 0.624474
## CAR_TYPEPanel Truck            0.1494173654  0.1139458070   1.311 0.189924
## CAR_TYPEPickup                 0.0349327315  0.0675504290   0.517 0.605126
## CAR_TYPESports Car             0.0361919156  0.0855726402   0.423 0.672391
## CAR_TYPESUV                    0.0208939265  0.0763594335   0.274 0.784404
## CAR_TYPEVan                    0.0074392884  0.0874055025   0.085 0.932182
## RED_CARyes                    -0.0045825782  0.0562131658  -0.082 0.935036
## OLDCLAIM                       0.0000023880  0.0000028221   0.846 0.397581
## CLM_FREQ                      -0.1045802790  0.1830292008  -0.571 0.567810
## REVOKEDYes                    -0.0892712012  0.0605611729  -1.474 0.140640
## MVR_PTS                        0.0023772773  0.0443129481   0.054 0.957222
## CAR_AGE                       -0.0544132053  0.0453975557  -1.199 0.230845
## URBANICITYHighly Urban/ Urban  0.0772653177  0.0849519007   0.910 0.363199
## CAR_AGE_BRAND_NEW_FLAG        -0.0359928141  0.2074810297  -0.173 0.862298
## CLM_FREQ_ZERO                  0.0957630861  0.7331820340   0.131 0.896096
## HOME_VAL_ZERO                 -0.1273463015  0.0874803662  -1.456 0.145648
## MVR_PTS_ZERO                  -0.0457748403  0.3164487673  -0.145 0.885002
## YOJ_ZERO                      -0.2234081327  0.1221966902  -1.828 0.067676 .
## I(BLUEBOOK^0.5)                0.0119039936  0.0032168563   3.701 0.000222 ***
## I(MVR_PTS^0.33)                0.0047714435  0.3206990707   0.015 0.988131
## I(CAR_AGE^0.5)                 0.2262671470  0.2857589599   0.792 0.428576
## I(CLM_FREQ^0.33)               0.2219460607  0.8880947239   0.250 0.802684
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8008 on 1786 degrees of freedom
## Multiple R-squared:  0.04176,    Adjusted R-squared:  0.01708
## F-statistic: 1.692 on 46 and 1786 DF,  p-value: 0.002745
```

```
##
## Call:
```

```
## lm(formula = log(TARGET_AMT) ~ YOJ + SEX + EDUCATION + BLUEBOOK +
##     CLM_FREQ + CAR_AGE + MVR_PTS_ZERO + YOJ_ZERO + I(BLUEBOOK^0.5) +
##     I(CAR_AGE^0.5) + I(CLM_FREQ^0.33), data = partial_train_mv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0976 -0.4089  0.0282  0.3993  3.2463
##
## Coefficients:
##                         Estimate  Std. Error t value Pr(>|t|)
## (Intercept)           7.53136444  0.21459431  35.096  < 2e-16 ***
## YOJ                  -0.01336321  0.00722113  -1.851 0.064394 .
## SEXM                  0.07397870  0.03802748   1.945 0.051881 .
## EDUCATIONBachelors   -0.08224430  0.06367141  -1.292 0.196626
## EDUCATIONHigh School -0.00004578  0.05396661  -0.001 0.999323
## EDUCATIONMasters      0.14096860  0.08416967   1.675 0.094143 .
## EDUCATIONPhD          0.23262536  0.10614081   2.192 0.028530 *
## BLUEBOOK             -0.00003438  0.00001219  -2.820 0.004853 **
## CLM_FREQ             -0.08574624  0.04135352  -2.073 0.038267 *
## CAR_AGE              -0.05663411  0.01992781  -2.842 0.004534 **
## MVR_PTS_ZERO         -0.06782767  0.04106551  -1.652 0.098770 .
## YOJ_ZERO             -0.20477123  0.10152579  -2.017 0.043849 *
## I(BLUEBOOK^0.5)       0.01027662  0.00292447   3.514 0.000452 ***
## I(CAR_AGE^0.5)        0.25065424  0.09251588   2.709 0.006806 **
## I(CLM_FREQ^0.33)      0.13661040  0.08226154   1.661 0.096949 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7977 on 1818 degrees of freedom
## Multiple R-squared:  0.03193,    Adjusted R-squared:  0.02447
## F-statistic: 4.283 on 14 and 1818 DF,  p-value: 0.0000001629
```

As with Model 5, while Model 6 results in a statistically significant p-value, the Adjusted R^2 suggests that Model 6 has little predictive power over our target variable.

Model 7 - AIC Stepwise model selection of Weighted Least Squares Multivariate Linear Regression

```
##
## Call:
## lm(formula = TARGET_AMT ~ KIDSDRIV + AGE + HOMEKIDS + YOJ + INCOME +
##     PARENT1 + HOME_VAL + MSTATUS + SEX + EDUCATION + JOB + TRAVTIME +
##     CAR_USE + BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM +
##     CLM_FREQ + REVOKED + MVR_PTS + CAR_AGE + URBANICITY + CAR_AGE_BRAND_NEW_FLAG +
##     CLM_FREQ_ZERO + HOME_VAL_ZERO + MVR_PTS_ZERO + YOJ_ZERO,
##     data = partial_train_mv, weights = 1/resid_sq)
##
## Weighted Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0496 -0.9926 -0.9600  0.9545  2.3695
##
## Coefficients:
##                                     Estimate    Std. Error t value Pr(>|t|)
```

```
## (Intercept)                      4873.0842551    84.3661857  57.761  < 2e-16 ***
## KIDSDRIV                         -111.9602346    15.7744441  -7.098 1.82e-12 ***
## AGE                                40.7258174     1.1986811  33.976  < 2e-16 ***
## HOMEKIDS                          392.4311076    10.6257129  36.932  < 2e-16 ***
## YOJ                               -39.4394157     3.4476425 -11.440  < 2e-16 ***
## INCOME                             -0.0041595     0.0005469  -7.605 4.56e-14 ***
## PARENT1Yes                       -125.3659963    35.3598233  -3.545 0.000402 ***
## HOME_VAL                           -0.0013641     0.0002593  -5.260 1.61e-07 ***
## MSTATUSYes                      -1166.9679554    30.4684593 -38.301  < 2e-16 ***
## SEXM                             1430.6332253    39.5380021  36.184  < 2e-16 ***
## EDUCATIONBachelors                142.8985715    32.9140776   4.342 1.49e-05 ***
## EDUCATIONHigh School             -173.4341133    31.8476691  -5.446 5.87e-08 ***
## EDUCATIONMasters                 1710.9516909    58.3192487  29.338  < 2e-16 ***
## EDUCATIONPhD                     2830.5017131    65.9090520  42.946  < 2e-16 ***
## JOBClerical                       -38.0882886    29.9556694  -1.271 0.203720
## JOBDoctor                       -2277.2474977   116.2062092 -19.597  < 2e-16 ***
## JOBHome Maker                    -576.1840898    55.0925830 -10.458  < 2e-16 ***
## JOBLawyer                          74.9249751    85.7254013   0.874 0.382229
## JOBManager                      -1117.8182762    60.6916068 -18.418  < 2e-16 ***
## JOBProfessional                  1083.1409584    39.8708612  27.166  < 2e-16 ***
## JOBStudent                        164.5639266    29.9942834   5.487 4.68e-08 ***
## JOBUnknown                        108.7377112    81.6214100   1.332 0.182957
## TRAVTIME                            1.0390187     0.5918803   1.755 0.079353 .
## CAR_USEPrivate                   -592.9245135    27.3582063 -21.673  < 2e-16 ***
## BLUEBOOK                            0.1032031     0.0022401  46.071  < 2e-16 ***
## TIF                               -11.1704211     2.1466194  -5.204 2.18e-07 ***
## CAR_TYPEPanel Truck             -1143.7951785    53.1445963 -21.522  < 2e-16 ***
## CAR_TYPEPickup                    -46.3061317    43.8658072  -1.056 0.291279
## CAR_TYPESports Car               1114.2649673    57.9203209  19.238  < 2e-16 ***
## CAR_TYPESUV                       781.1395208    52.6566308  14.835  < 2e-16 ***
## CAR_TYPEVan                       254.2781843    54.3452338   4.679 3.10e-06 ***
## RED_CARyes                       -477.8713007    28.9438369 -16.510  < 2e-16 ***
## OLDCLAIM                            0.0210646     0.0018387  11.456  < 2e-16 ***
## CLM_FREQ                         -254.6347971    12.8456531 -19.823  < 2e-16 ***
## REVOKEDYes                      -1405.7877024    32.6326465 -43.079  < 2e-16 ***
## MVR_PTS                            54.0500296     5.0069287  10.795  < 2e-16 ***
## CAR_AGE                          -210.3342257     3.3401131 -62.972  < 2e-16 ***
## URBANICITYHighly Urban/ Urban     736.9820612    30.2400774  24.371  < 2e-16 ***
## CAR_AGE_BRAND_NEW_FLAG          -1549.7110334    29.0037684 -53.431  < 2e-16 ***
## CLM_FREQ_ZERO                    -179.8828706    40.1730267  -4.478 8.02e-06 ***
## HOME_VAL_ZERO                   -1003.8803414    54.6375143 -18.373  < 2e-16 ***
## MVR_PTS_ZERO                      -68.5349793    27.7667595  -2.468 0.013671 *
## YOJ_ZERO                         -965.4437850    50.6080486 -19.077  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9939 on 1790 degrees of freedom
## Multiple R-squared:  0.9935, Adjusted R-squared:  0.9933
## F-statistic:  6493 on 42 and 1790 DF,  p-value: < 2.2e-16
```

Our Adjusted R^2 value here is 99% which, at face value, appears like a miraculous improvement upon our earlier modeling attempts. However, the weighted least squares R^2 value cannot be interpreted in the same way that the unweighted R^2 from ordinary least squares linear regression models. Regardless, based on the F-statistic and the p-value, it looks like the model is extremely significant as well.

In looking at these coefficients we note some things that we wouldn't have expected:

- The older you are the more expensive your wreck will be (perhaps older people drive more expensive cars)
- The more kids you have at home, the more expensive your wreck will be
- Being a man significantly increases the cost of your wreck (maybe correlated with the cost of the car being driven?)
- Masters and PhD level individuals have more expensive wrecks (maybe because they drive more expensive cars?)
- The longer your travel time, the less expensive the wreck will cost
- Bluebook value does not seem to be as a large a factor as expected
- Red cars actually would cost less in an accident (opposite of urban legend)

Model evaluation and selection

Binary logistic regression

In looking at our binary logistic regression models, we'll evaluate our binary logistic regression model (Model 1), the stepwise binary logistic regression model (Model 2), as well as the Random Forest model (Model 3) as a benchmark for accuracy.

Evaluate Model 1 Binary Logistic Model

```
## [1] 0.782502
```

```
##     predicted
## true   0   1
##    0 823  78
##    1 188 134
```

We see the accuracy of this model is 79%. Our precision is 93% and our sensitivity is 82%.

Evaluate Model 2 Stepwise Binary Logistic Model

```
## [1] 0.7800491
```

```
##     predicted
## true   0   1
##    0 823  78
##    1 191 131
```

Our stepwise model produces *VERY* similar results to our previous model. Results vary only slightly.

Evaluate Model 3 Random Forest

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
```

```
##          0 837 195
##          1  64 127
##
##                Accuracy : 0.7882
##                  95% CI : (0.7642, 0.8108)
##     No Information Rate : 0.7367
##     P-Value [Acc > NIR] : 1.687e-05
##
##                   Kappa : 0.372
##
##  Mcnemar's Test P-Value : 6.594e-16
##
##             Sensitivity : 0.9290
##             Specificity : 0.3944
##          Pos Pred Value : 0.8110
##          Neg Pred Value : 0.6649
##              Prevalence : 0.7367
##          Detection Rate : 0.6844
##    Detection Prevalence : 0.8438
##       Balanced Accuracy : 0.6617
##
##        'Positive' Class : 0
##
```

Here we see our random forest model is in line with our other models as far as accuracy, however, we see a trade off in terms of precision and sensitivity. Our accuracy is only **0.7882257%** whereas our sensitivity is now **0.9289678%**.

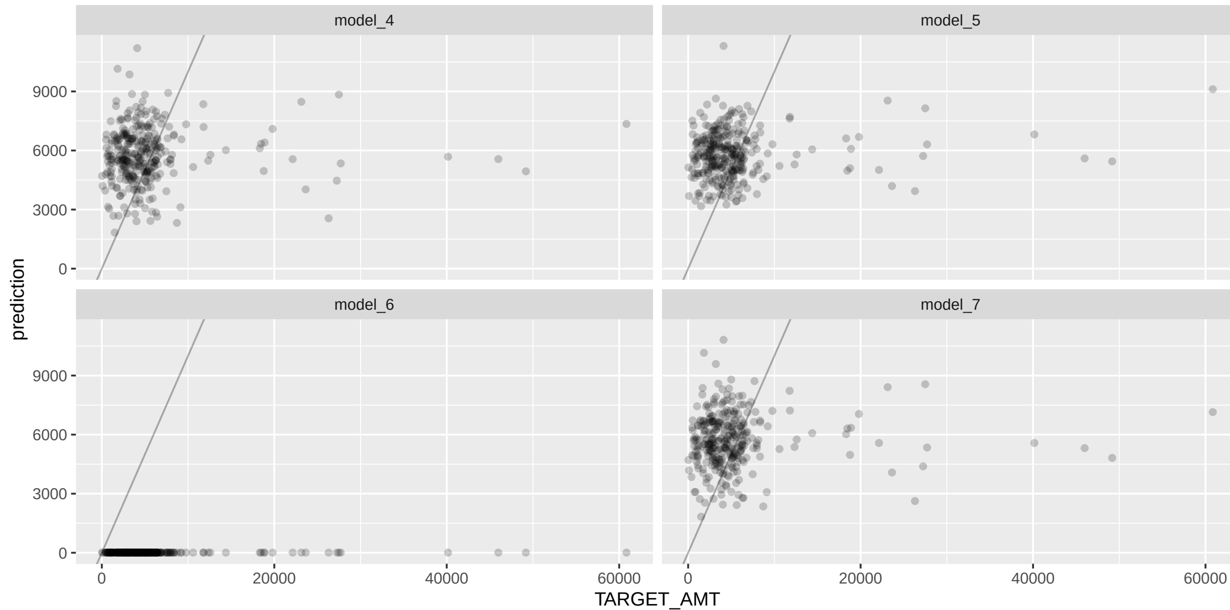
For our current model, we'd probably be most interested in sensitivity since we're concerned with identifying positive outcomes and the cost of a false-positive is low. Were we looking for the most accurate model, we'd move forward with the random forest model, however, for this assignment we'll continue forward with our most accurate binary logistic regression model.

Summary for Binary Logistic Regression Models

We can see that out of all the models the Random Forest and the Binomial Model using Stepwise were practically tied, with the Random Forest being slightly more accurate. As mentioned previously, we'll move forward with the stepwise logistic regression model.

Multivariate Linear Regression

Multivariate linear regression approaches resulted in statistically significant fits to the data, however, the model predictions are have room for improvement. The figure below shows each multivariate model's predictions plotted as a function of response variable. Ideally, this plot would result in an approximately linear agreement. However, we see a lot of deviation which suggests our multivariate models we deployed do not make accurate predictions of the numeric target variable `TARGET_AMT`




```
## RMSE
##  Model 4:  6621.119
##  Model 5:  6498.003
##  Model 6:  8466.536
##  Model 7:  6623.416
```

Comparing the RMSE of each of the models on the holdout validation dataset, it looks like Model 5 has the smallest errors.

Model Selection Summary

In our analysis, we find that multivariate linear regression modeling approaches have a lot of room for improvement towards predictions of the numeric target feature `TARGET_AMT`. However, the binary logistic regression models were able to give predictions of the binary response variable `TARGET_FLAG` that are reasonably accurate. Moving forward, we will use the Stepwise Binary Logistic Model (Model 2) to predict the probability that a person will crash their car (`TARGET_FLAG`) and the Stepwise Multivariate Linear Regression (Model 5) to give prediction estimates for the amount of money it will cost if a person crashes their car (`TARGET_AMT`).

Make Prediction on Test Data with Best Model

Logistic regression predictions: - predictions on `TARGET_FLAG`, the probability that a person will crash their car

```
test <- final_df %>% filter(dataset == 'test') %>% dplyr::select(-dataset)

logistic_binary_final <- predict(binary.mdl.w.step, test, type = "response")
head(logistic_binary_final)
```

```
##         1         2         3         4         5         6
## 0.1323629 0.2677470 0.1328139 0.3139109 0.1323659 0.2772592
```

Multivariate Regression Predictions: - predictions on `TARGET_AMT`, amount of money it will cost if a person crashes their car

```
MVPred <- predict(lm2, newdata = test)
head( MVPred)
```

```
##        1        2        3        4        5        6
## 7153.636 6733.137 5255.977 5595.512 6941.973 6853.909
```

Reference Section:

- Practical Guide to Logistic Regression Analysis in R

Appendix: R Statistical Code

Dependancies

```
# Libraries and Options
knitr::opts_chunk$set(echo = F, warning = F, message = F, eval = T,
                      fig.height = 5, fig.width = 10)

library(knitr)
library(skimr)
library(visdat)
library(inspectdf)
library(corrplot)
library(scales)
library(tidyverse)
library(tidyr)
library(bestglm)
library(pROC)
library(car)
library(ggcorrplot)
library(mice)
library(caret)
library(plyr)
library(dplyr)
library(MASS)
library(zoo)


options(scipen = 9)
set.seed(123)

boxplot_depend_vs_independ <- function(df_train, target_name) {

  train_int_names <- df_train %>% select_if(is.numeric)

  int_names <- names(train_int_names)
```

```r
  myGlist <- vector('list', length(int_names))

  names(myGlist) <- int_names

  for (i in int_names) {

   myGlist[[i]] <-
       ggplot(df_train, aes_string(x = target_name, y = i)) +
        geom_boxplot(color = 'steelblue', outlier.color = 'firebrick',
                     outlier.alpha = 0.35) +
        labs(title = paste0(i,' vs target'), y = i, x= 'target') +
        theme_minimal() +
        theme(
          plot.title = element_text(hjust = 0.45),
          panel.grid.major.y =  element_line(color = "grey",
                                           linetype = "dashed"),
          panel.grid.major.x = element_blank(),
          panel.grid.minor.y = element_blank(),
          panel.grid.minor.x = element_blank(),
          axis.ticks.x = element_line(color = "grey")
        )

      }

    myGlist <- within(myGlist, rm(target_name))
    gridExtra::grid.arrange(grobs = myGlist, ncol = 3)
}


plot_corr_matrix <- function(dataframe, significance_threshold){
  title <- paste0('Correlation Matrix for significance > ',
                significance_threshold)

  df_cor <- dataframe %>% mutate_if(is.character, as.factor)

  df_cor <- df_cor %>% mutate_if(is.factor, as.numeric)
  #run a correlation and drop the insignificant ones
  corr <- cor(df_cor)
  #prepare to drop duplicates and correlations of 1
  corr[lower.tri(corr,diag=TRUE)] <- NA
  #drop perfect correlations
  corr[corr == 1] <- NA
  #turn into a 3-column table
  corr <- as.data.frame(as.table(corr))
  #remove the NA values from above
  corr <- na.omit(corr)
  #select significant values
  corr <- subset(corr, abs(Freq) > significance_threshold)
  #sort by highest correlation
  corr <- corr[order(-abs(corr$Freq)),]
  #print table
  # print(corr)
  #turn corr back into matrix in order to plot with corrplot
```

```r
  mtx_corr <- reshape2::acast(corr, Var1~Var2, value.var="Freq")

  #plot correlations visually
  corrplot(mtx_corr,
           title=title,
           mar=c(0,0,1,0),
           method='color',
           tl.col="black",
           na.label= " ",
           addCoef.col = 'black',
           number.cex = .9)
}
```

Importing Data

```r
test_URL <- paste0('https://raw.githubusercontent.com/AngelClaudio/',
                   'data-sources/master/csv/insurance-evaluation-data.csv')

train_URL <- paste0('https://raw.githubusercontent.com/AngelClaudio/',
                    'data-sources/master/csv/insurance_training_data.csv')

train <- readr::read_csv(train_URL)
test <- readr::read_csv(test_URL)
```

Data Exploration

```r
glimpse(train) #basic visualization
dim(train %>% select_if(is.numeric))[2] #number of numeric features
dim(train %>% select_if(is.character))[2] #number categorical features
summary(train) #basic summary
colSums(is.na(train)) #missingness by column
#visualization to find patterns in missingness
VIM::aggr(train, col=c('green','red'), numbers=T, sortVars=T,
          cex.axis = .7,
          ylab=c("Proportion of Data", "Combinations and Percentiles"))
```

Data Preparation

```r
train$dataset <- 'train'
test$dataset <- 'test'
final_df <- rbind(train, test) #combine test/train to preprocess the same
final_df <-  dplyr::select(final_df, -c('INDEX')) #drop INDEX feature
#change the following from character to numeric
final_df  <- dplyr::mutate(final_df, INCOME = as.numeric(gsub('[$,]', '', INCOME)),
                            HOME_VAL = as.numeric(gsub('[$,]', '', HOME_VAL)),
                            BLUEBOOK = as.numeric(gsub('[$,]', '', BLUEBOOK)),
                            OLDCLAIM = as.numeric(gsub('[$,]', '', OLDCLAIM)))
```

```r
#transform from character to factor:
final_df <- dplyr::mutate(final_df,
                          MSTATUS = as.factor(str_remove(MSTATUS, "^z_")),
                          SEX = as.factor(str_remove(SEX, "^z_")),
                          # <HIgh School not a typo, means less than HS
                          EDUCATION = as.factor(str_remove(EDUCATION, "^z_")),
                          JOB = as.factor(str_remove(JOB, "^z_")),
                          CAR_TYPE = as.factor(str_remove(CAR_TYPE, "^z_")),
                          URBANICITY = as.factor(str_remove(URBANICITY, "^z_")),
                          CAR_USE = as.factor(CAR_USE),
                          REVOKED = as.factor(REVOKED),
                          PARENT1 = as.factor(PARENT1),
                          RED_CAR = as.factor(RED_CAR),
                          TARGET_FLAG = as.factor(TARGET_FLAG))
#na.approx from the zoo library to perform linear interpolation on NA values
final_df <- final_df %>% mutate_at(vars(c("CAR_AGE", "YOJ", "AGE", "INCOME",
                                    "HOME_VAL")),
                             ~ifelse(is.na(.), na.approx(.), .))
# impute NAs in job
final_df$JOB <- as.character(final_df$JOB)
final_df$JOB[is.na(final_df$JOB)] <- "Unknown"
final_df$JOB <- as.factor(final_df$JOB)
#visualize missingness post imputation
sapply(final_df, function(x) sum(is.na(x)))
VIM::aggr(final_df, col=c('green','red'), numbers=T, sortVars=T,
          cex.axis = .7,
          ylab=c("Proportion of Data", "Combinations and Percentiles"))
# unbind data
train <- dplyr::select(dplyr::filter(final_df, dataset == 'train'),
                       -c('dataset'))
test <- dplyr::select(dplyr::filter(final_df, dataset == 'test'),
                       -c('dataset'))
# factor analysis. visualize most common factors
inspectdf::inspect_imb(train) %>% show_plot()
# visualize a summary of common levels
inspectdf::inspect_num(train) %>% show_plot()
# boxplots of features x predictor
target_name <- 'TARGET_FLAG'
boxplot_depend_vs_independ(train, target_name)
#plot the correlation matrix
plot_corr_matrix(train, .2)
# feature engineering
# flag brand new cars
final_df$CAR_AGE <- ifelse(final_df$CAR_AGE < 1, 1, final_df$CAR_AGE)
final_df$CAR_AGE_BRAND_NEW_FLAG <- ifelse(final_df$CAR_AGE == 1, 1, 0)
# zero claims
final_df$CLM_FREQ_ZERO <- ifelse(final_df$CLM_FREQ == 0, 1, 0)
# Home Value
final_df$HOME_VAL_ZERO <- ifelse(final_df$HOME_VAL == 0, 1, 0)
# Motor Vehicle Record Points
final_df$MVR_PTS_ZERO <- ifelse(final_df$MVR_PTS == 0, 1, 0)
# no years on job
final_df$YOJ_ZERO <- ifelse(final_df$YOJ == 0, 1, 0)
```

Build Models

```r
# Prepare data
train_data <- final_df %>% filter(dataset == 'train') %>% dplyr::select(-dataset)
train_data$TARGET_AMT <- NULL
split <- caret::createDataPartition(train_data$TARGET_FLAG, p=0.85, list=FALSE)
partial_train <- train_data[split, ]
validation <- train_data[ -split, ]

# Model 1
binary.mdl <- glm(TARGET_FLAG~., family=binomial, data=partial_train)
summary(binary.mdl)

# Model 2
binary.mdl.w.step <- step(binary.mdl)
summary(binary.mdl.w.step)

# Model 3
rf <- randomForest::randomForest(TARGET_FLAG~., data=partial_train)

# Prepare data
train_mv_mdl <- final_df %>% filter(dataset == 'train') %>% dplyr::select(-dataset)
train_mv_mdl <- train_mv_mdl[train_mv_mdl$TARGET_FLAG == 1, ]
train_mv_mdl$TARGET_FLAG <- NULL
split_mv <- caret::createDataPartition(train_mv_mdl$TARGET_AMT, p=0.85, list = F)
partial_train_mv <- train_mv_mdl[split_mv, ]
validation_mv <- train_mv_mdl[-split_mv, ]

# Model 4
mv.mdl <- train(TARGET_AMT ~., data = partial_train_mv, method = "lm",
                 trControl = trainControl(method = "cv", number = 10,
                                          savePredictions = TRUE),
                tuneLength = 5, preProcess = c("center", "scale"))
summary(mv.mdl$finalModel)
lm1 <- mv.mdl

# Model 5
lm2_base <- lm(TARGET_AMT ~ ., data = partial_train_mv)
lm_step <- stepAIC(lm2_base, trace = F)
summary(lm_step)
lm2 <- lm_step

# Model 6
train_mv_mdl %>%
  select_if(is.numeric) %>%
  dplyr::mutate(row = row_number()) %>%
  tidyr::gather(field, val, -row) %>%
  dplyr::filter(val > 0) %>%
  tidyr::spread(field, val) %>%
  dplyr::select(-row, -CAR_AGE_BRAND_NEW_FLAG, -CLM_FREQ_ZERO, -HOME_VAL_ZERO, -MVR_PTS_ZERO, -YOJ_ZERO
  powerTransform()
lm3_base <- lm(log(TARGET_AMT) ~ . + I(BLUEBOOK^0.5) + I(MVR_PTS^.33) + I(CAR_AGE^.5) + I(CLM_FREQ^0.33
               , data = partial_train_mv)
```

```r
summary(lm3_base)
lm3_step <- stepAIC(lm3_base, trace = F, direction = 'backward')
summary(lm3_step)
lm3 <- lm3_step

# Model 7
lm4_base <- lm(TARGET_AMT ~ ., data = partial_train_mv)
resid_sq <- lm4_base$residuals^2
lm4_wls <- lm(TARGET_AMT ~ ., data = partial_train_mv, weights = 1/resid_sq)
summary(lm4_wls)
lm4_wls_step <- stepAIC(lm4_wls, trace = F)
summary(lm4_wls_step)
#
```

Model Selection

```r
#evaluate Model 1
y_hat_glm <- predict(binary.mdl, validation, type = "response")
y_hat_glm_binar <- (y_hat_glm>0.5)*1
mean(validation$TARGET_FLAG==y_hat_glm_binar)
(CM <- table(true= validation$TARGET_FLAG, predicted = y_hat_glm_binar))
#Evaluate Model 2
y_hat_glm <- predict(binary.mdl.w.step, validation, type = "response")
y_hat_glm_binar <- (y_hat_glm>0.5)*1
mean(validation$TARGET_FLAG==y_hat_glm_binar)
(CM <- table(true= validation$TARGET_FLAG, predicted = y_hat_glm_binar))
#Evaluate Model 3
val <- validation
val$pred <- predict(rf, val)
val$pred <- as.factor(val$pred)
confusionMatrix(val$pred, val$TARGET_FLAG)

# Evaluating Model 4-7
validation_mv$pred1 <- predict(lm1, newdata = validation_mv)
validation_mv$pred2 <- predict(lm2, newdata = validation_mv)
validation_mv$pred3 <- predict(lm3, newdata = validation_mv)
#validation_mv$pred4 <- predict(lm4, newdata = validation_mv)
validation_mv %>%
  dplyr::select(TARGET_AMT, pred1, pred2, pred3) %>%
  tidyr::gather(model, prediction, -TARGET_AMT) %>%
  ggplot(aes(x = TARGET_AMT, y = prediction)) +
  geom_point(alpha = .2) +
  geom_abline(slope = 1, intercept = 0, alpha = .3) +
  facet_wrap(~model)
validation_mv$resid1 <- with(validation_mv, pred1 - TARGET_AMT)
validation_mv$resid2 <- with(validation_mv, pred2 - TARGET_AMT)
validation_mv$resid3 <- with(validation_mv, pred3 - TARGET_AMT)
#validation_mv$resid4 <- with(validation_mv, pred4 - TARGET_AMT)
validation_mv %>%
  dplyr::select(TARGET_AMT, resid1, resid2, resid3) %>%
  tidyr::gather(model, residuals, -TARGET_AMT) %>%
  ggplot(aes(x = TARGET_AMT, y = residuals)) +
  geom_point(alpha = .2) +
  geom_hline(yintercept = 0, alpha = .3) +
```

```r
  facet_wrap(~model)

# RMSE calc
validation_mv$resid_4 <- with(validation_mv, model_4 - TARGET_AMT)
validation_mv$resid_5 <- with(validation_mv, model_5 - TARGET_AMT)
validation_mv$resid_6 <- with(validation_mv, model_6 - TARGET_AMT)
validation_mv$resid_7 <- with(validation_mv, model_7 - TARGET_AMT)

cat('RMSE\n',
    'Model 4: ', sqrt(sum(validation_mv$resid_4^2) / nrow(validation_mv)), '\n',
    'Model 5: ', sqrt(sum(validation_mv$resid_5^2) / nrow(validation_mv)), '\n',
    'Model 6: ', sqrt(sum(validation_mv$resid_6^2) / nrow(validation_mv)), '\n',
    'Model 7: ', sqrt(sum(validation_mv$resid_7^2) / nrow(validation_mv))
)
```

Model Predictions

```r
test <- final_df %>% filter(dataset == 'test') %>% dplyr::select(-dataset)
logistic_binary_final <- predict(binary.mdl.w.step, test, type = "response")
logistic_binary_final
predict(lm3, newdata = test)
```