DATA 621 Project 4

Michael Muller

April 22, 2018

Overview

The task

Finalize two models using the insurance_training_data.csv

|Binary logistic regression model| *predicting whether or not a person will crash their car*

|Multivariate linear regression model| *predicting the cost of crashing a car*

The restraints

Only data from the training_set may be used

|The training_set contains| *8161 raw observations*

|The training_set contains| *26 features (including index and targets)*

1. DATA EXPLORATION

Taking a look into our dataset

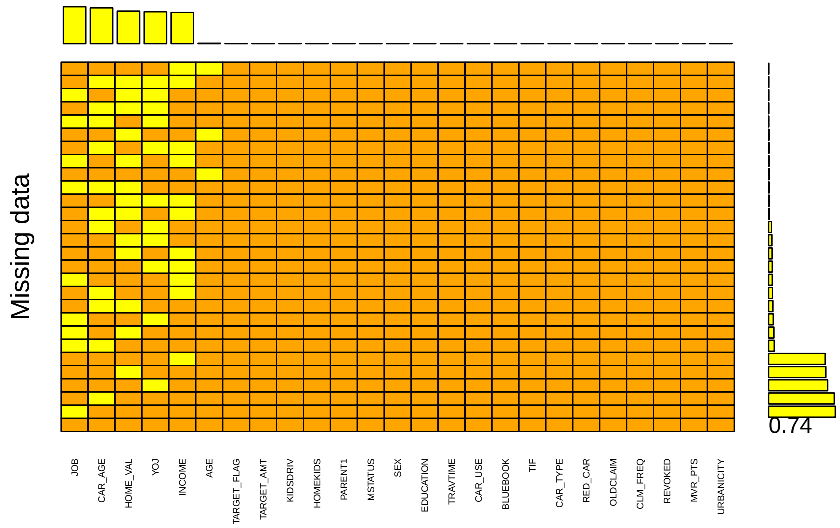
INDEX	TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL	MSTATUS	SEX	EDUCATION	JOB	TRAVTIME
1	0	0	0	60	0	11	$67,349	No	$0	z_No	M	PhD	Professional	14
2	0	0	0	43	0	11	$91,449	No	$257,252	z_No	M	z_High School	z_Blue Collar	22
4	0	0	0	35	1	10	$16,039	No	$124,191	Yes	z_F	z_High School	Clerical	5
5	0	0	0	51	0	14		No	$306,251	Yes	M	<High School	z_Blue Collar	32
6	0	0	0	50	0	NA	$114,986	No	$243,925	Yes	z_F	PhD	Doctor	36
7	1	2946	0	34	1	12	$125,301	Yes	$0	z_No	z_F	Bachelors	z_Blue Collar	46

CAR_USE	BLUEBOOK	TIF	CAR_TYPE	RED_CAR	OLDCLAIM	CLM_FREQ	REVOKED	MVR_PTS	CAR_AGE	URBANICITY
Private	$14,230	11	Minivan	yes	$4,461	2	No	3	18	Highly Urban/ Urban
Commercial	$14,940	1	Minivan	yes	$0	0	No	0	1	Highly Urban/ Urban
Private	$4,010	4	z_SUV	no	$38,690	2	No	3	10	Highly Urban/ Urban
Private	$15,440	7	Minivan	yes	$0	0	No	0	6	Highly Urban/ Urban
Private	$18,000	1	z_SUV	no	$19,217	2	Yes	3	17	Highly Urban/ Urban
Commercial	$17,430	1	Sports Car	no	$0	0	No	0	7	Highly Urban/ Urban

This is no good; we need to tidy this data, and create a method to tidy further datasets of the same features. Features such as CAR_USE (with two possible values) should be converted to binary format, and miscellaneous text such as '$' and 'z_' should be filtered out of any value for easier analysis.

Lets check out the data's sparsity using the MICE package in R.
Interpretting the figure below; we have

```
## Warning in plot.aggr(res, ...): not enough vertical space to display
## frequencies (too many combinations)
```




```
##
##  Variables sorted by number of missings:
##     Variable Count
##          JOB   526
##      CAR_AGE   510
##     HOME_VAL   464
##          YOJ   454
##       INCOME   445
##          AGE     6
##  TARGET_FLAG     0
##   TARGET_AMT     0
##     KIDSDRIV     0
##     HOMEKIDS     0
##      PARENT1     0
##      MSTATUS     0
##          SEX     0
##    EDUCATION     0
##     TRAVTIME     0
##      CAR_USE     0
##     BLUEBOOK     0
##          TIF     0
##     CAR_TYPE     0
##      RED_CAR     0
##     OLDCLAIM     0
##     CLM_FREQ     0
##      REVOKED     0
##      MVR_PTS     0
##   URBANICITY     0
```

Awesome! No missing target data. Most missing data is numeric so we can use data imputation to give us a fuller set of data to work with. More on that during DATA PREPERATION.

Lets move on to the distributions of our numeric data, checking boxplots and histograms.

Figure 1 : Boxplots

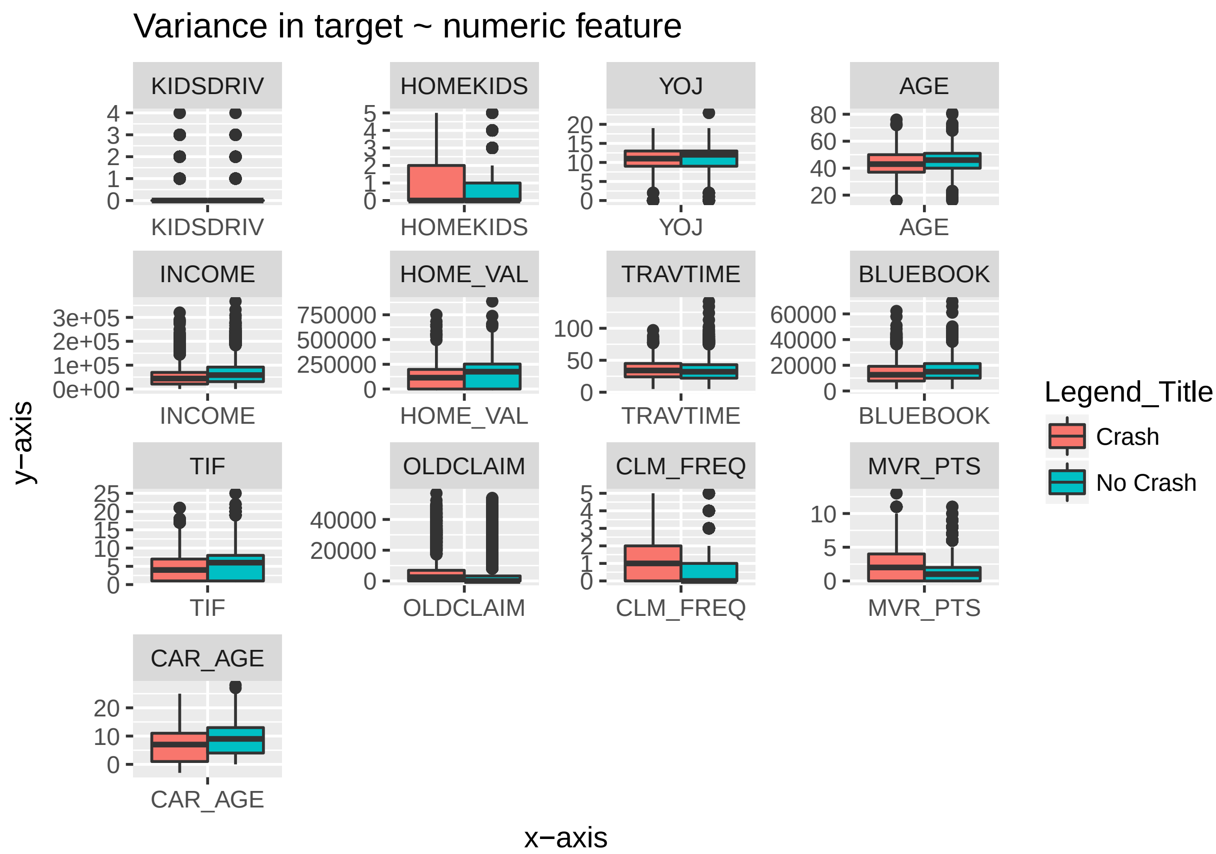


Figure 2 : Histograms

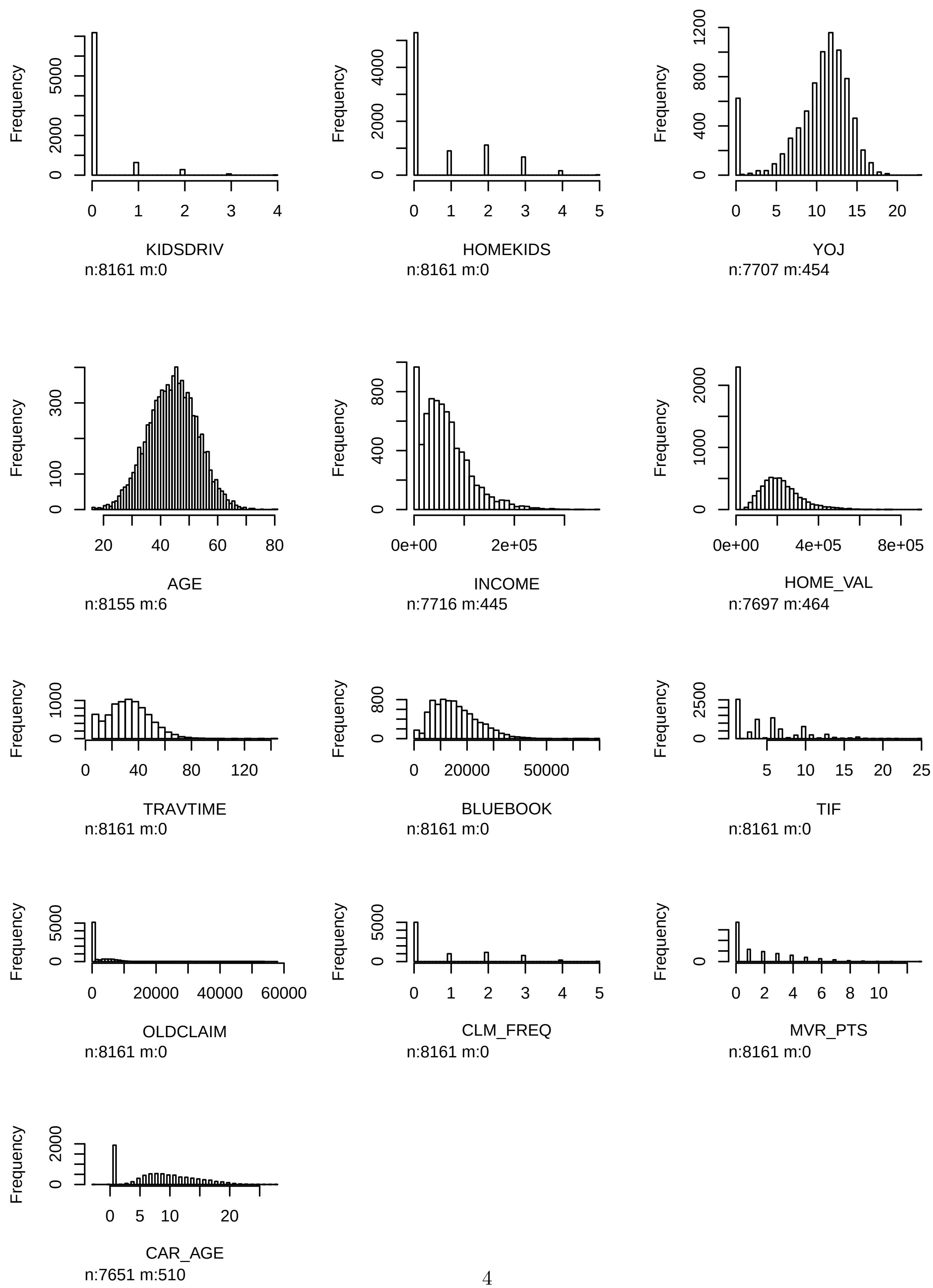
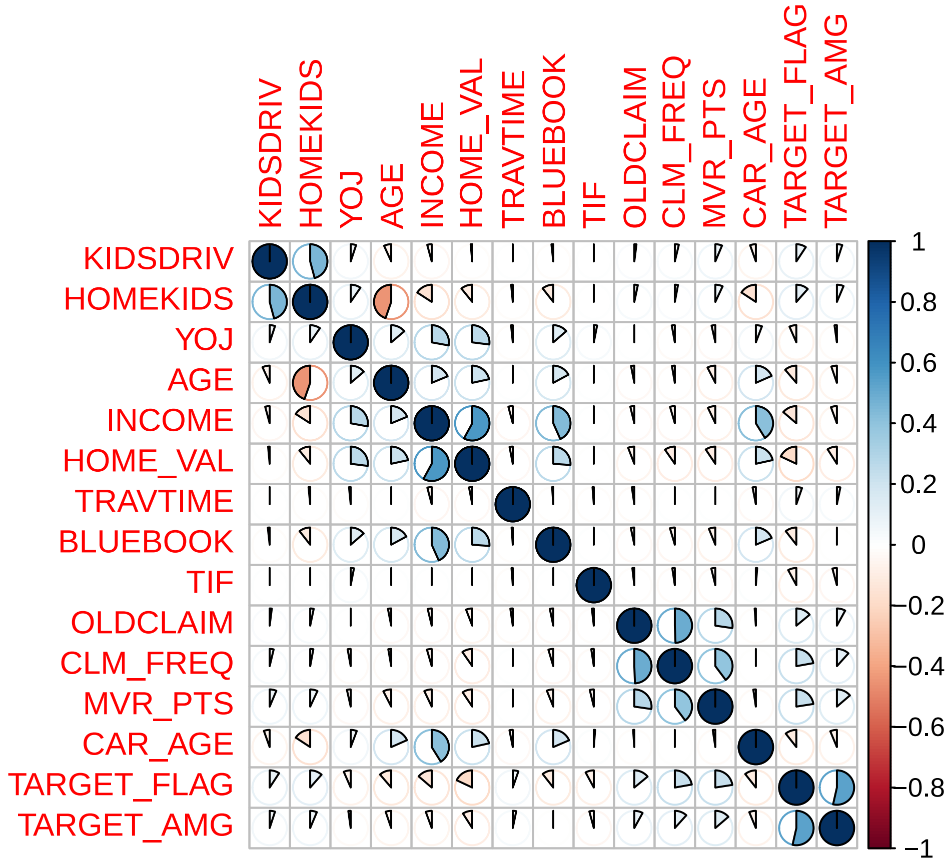


Figure 3 : Correlation matrix plot



The following figures present both good and bad information.

|The Good| We have near normal distributions for bluebook, travtime, YOJ, age, and INCOME. Our boxplots show some even TARGET prediction spreads, and the correlation matrix shows no issues of multi-colinearity.

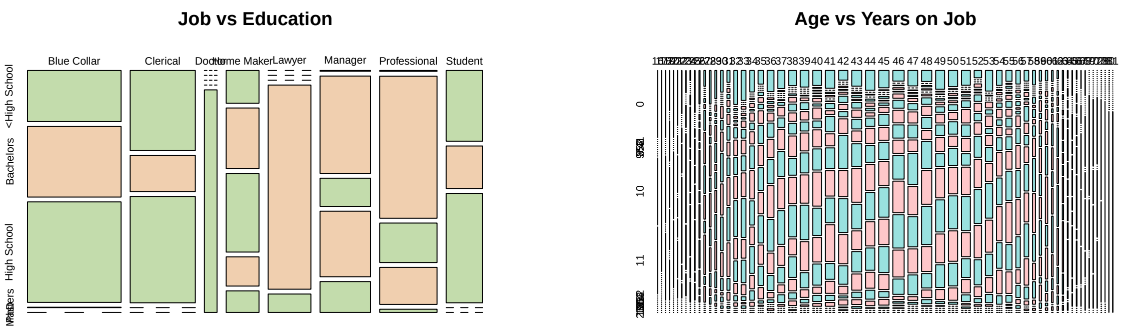
We may want to take the log of BLUEBOOK,TRAVTIME, YOJ for more normal distributions

|The Bad| We have high frequency 0 values in most of our distributions and CLM_FREQ boxplot shows we may need to alter that variable.

Which means we may need to treat many of these variables as factors, or make new features entirely.

Lets look at a few mosaic plots to see the interaction between a few variables for input on possible imputations.

Figure 4 : Mosaic plot



We can see two trends from the two mosaic plots (Ones a bit harder to see).

|Trend 1| Higher education means higher ranking jobs. However there is too much variation to make any other assumption that a lawyer has a bachelors which already accounted for.

|Trend 2| Aside from the 50th percentile area; YOJ goes up with age

Unfortunately, Trend 2 is too weak to base imputation on. . . we'll use traditional methods instead

2. DATA PREPERATION

Outline of data preperation

Issue | Remedy

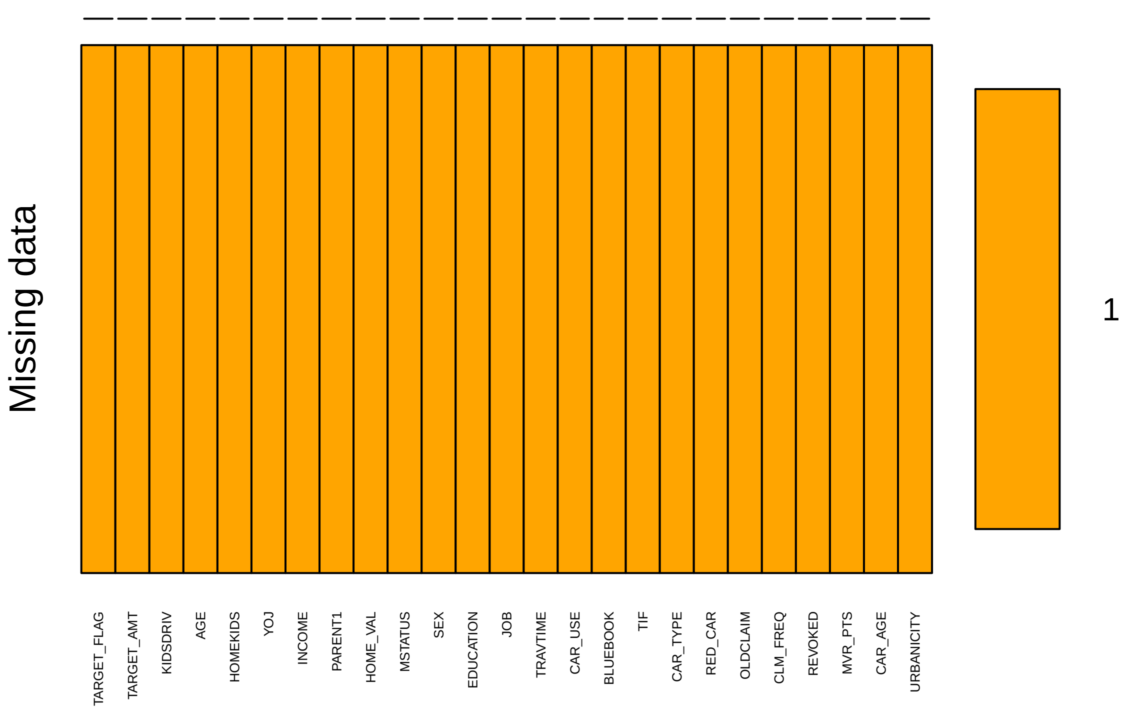
Missing Values | Imputation.

Many distributions are bimodal | Make binary features from existing features.

Too many features, parsimony issue | Restrict to maximum of 10 predictors per model.

Many distributions can be normalized | Take log

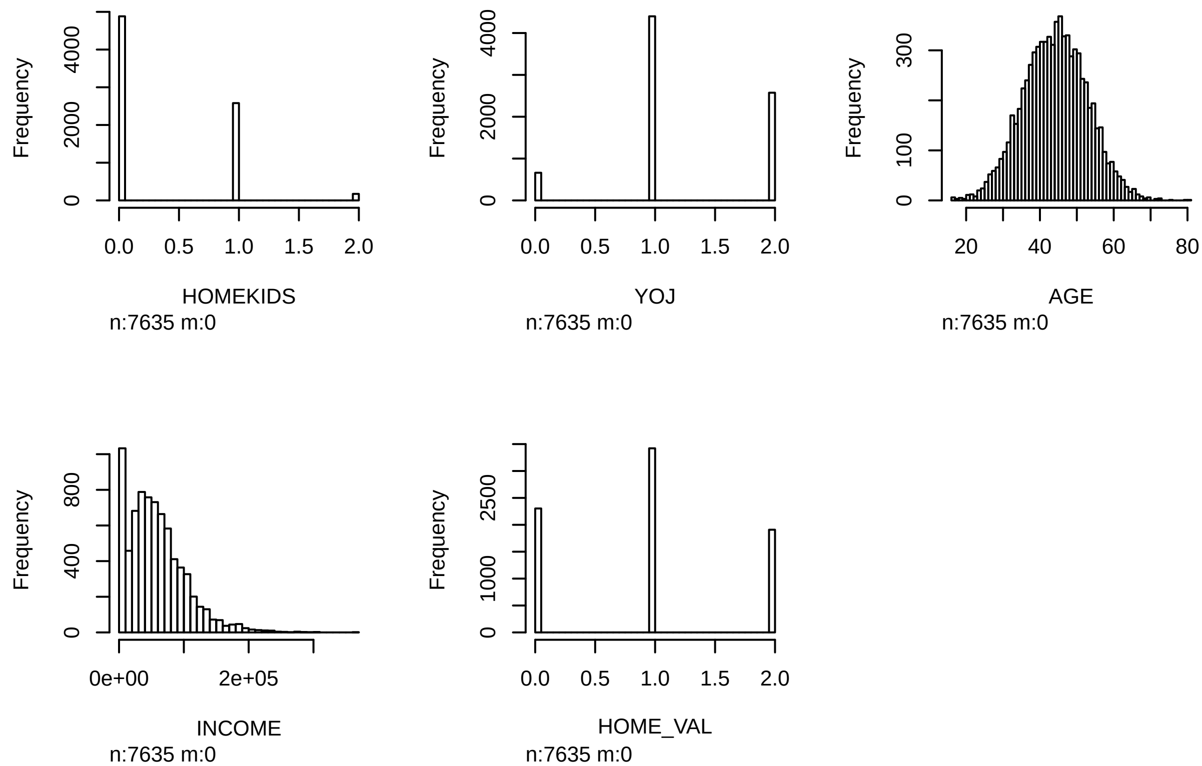
```
##
##  iter imp variable
##   1   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   1   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   2   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   3   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   4   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   1  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
##   5   2  AGE  YOJ  INCOME  HOME_VAL  CAR_AGE
```

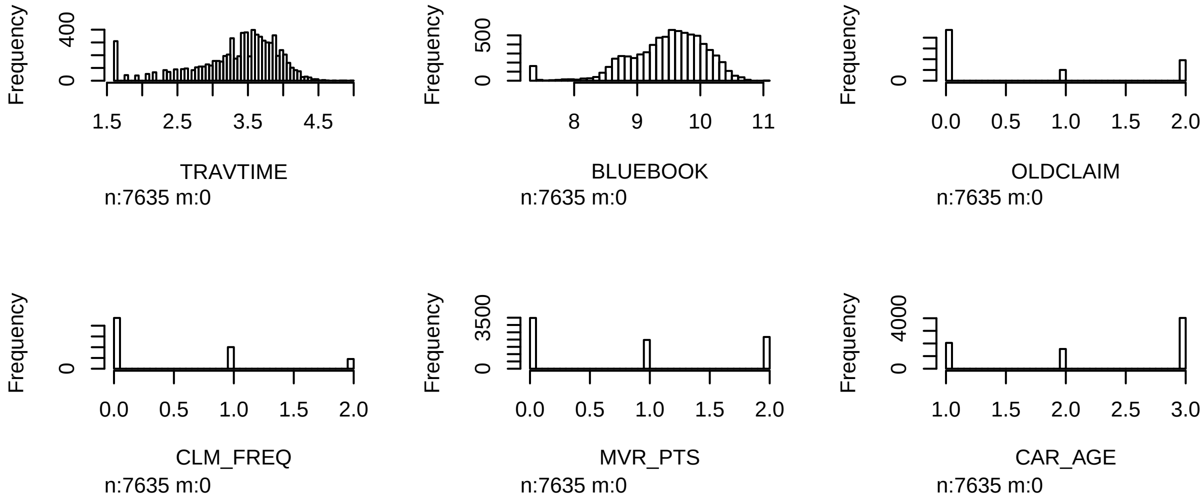



```
##
##  Variables sorted by number of missings:
##     Variable Count
##  TARGET_FLAG     0
##   TARGET_AMT     0
##     KIDSDRIV     0
##          AGE     0
##     HOMEKIDS     0
##          YOJ     0
##       INCOME     0
##      PARENT1     0
##     HOME_VAL     0
##      MSTATUS     0
##          SEX     0
##    EDUCATION     0
##          JOB     0
##     TRAVTIME     0
##      CAR_USE     0
##     BLUEBOOK     0
##          TIF     0
##     CAR_TYPE     0
##      RED_CAR     0
##     OLDCLAIM     0
##     CLM_FREQ     0
##      REVOKED     0
##      MVR_PTS     0
```

```
##      CAR_AGE     0
##   URBANICITY     0
```

Good, no missing data. Lets go on categorizing our oddly distributed features





3. BUILD MODELS

```
##
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = m1d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5509  -0.7069  -0.3866   0.6196   3.1661
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.257e+00  2.857e-01 -11.403  < 2e-16 ***
## KIDSDRIV              3.637e-01  6.270e-02   5.800 6.63e-09 ***
## AGE                  -1.892e-03  4.165e-03  -0.454 0.649642
## HOMEKIDS              4.293e-02  3.813e-02   1.126 0.260274
## YOJ                  -1.128e-02  8.568e-03  -1.316 0.188125
## INCOME               -4.209e-06  1.264e-06  -3.329 0.000872 ***
## PARENT1               3.467e-01  1.135e-01   3.055 0.002250 **
## HOME_VAL             -1.320e-06  3.841e-07  -3.437 0.000588 ***
## MSTATUS              -4.940e-01  8.967e-02  -5.509 3.60e-08 ***
## SEX                   1.280e-01  1.164e-01   1.100 0.271545
## EDUCATIONBachelors   -3.430e-01  1.165e-01  -2.943 0.003247 **
## EDUCATIONHigh School  2.289e-02  9.532e-02   0.240 0.810204
```

```
## EDUCATIONMasters     -2.827e-01  1.842e-01  -1.535 0.124803
## EDUCATIONPhD          8.749e-02  2.316e-01   0.378 0.705630
## JOBClerical           8.790e-02  1.077e-01   0.816 0.414342
## JOBDoctor            -8.927e-01  3.021e-01  -2.955 0.003127 **
## JOBHome Maker        -1.335e-01  1.551e-01  -0.861 0.389457
## JOBLawyer            -1.432e-01  1.896e-01  -0.755 0.450229
## JOBManager           -8.512e-01  1.403e-01  -6.067 1.31e-09 ***
## JOBProfessional      -1.297e-01  1.205e-01  -1.076 0.281810
## JOBStudent           -1.433e-01  1.314e-01  -1.090 0.275580
## TRAVTIME              1.464e-02  1.950e-03   7.511 5.87e-14 ***
## CAR_USE               7.772e-01  9.356e-02   8.308  < 2e-16 ***
## BLUEBOOK             -2.114e-05  5.468e-06  -3.865 0.000111 ***
## TIF                  -5.469e-02  7.644e-03  -7.154 8.45e-13 ***
## CAR_TYPEPanel Truck   5.973e-01  1.756e-01   3.403 0.000668 ***
## CAR_TYPEPickup        5.825e-01  1.025e-01   5.682 1.33e-08 ***
## CAR_TYPESports Car    1.031e+00  1.315e-01   7.842 4.44e-15 ***
## CAR_TYPESUV           7.726e-01  1.130e-01   6.839 8.00e-12 ***
## CAR_TYPEVan           5.702e-01  1.328e-01   4.294 1.75e-05 ***
## RED_CAR              -1.020e-01  9.238e-02  -1.104 0.269477
## OLDCLAIM             -1.442e-05  4.104e-06  -3.513 0.000444 ***
## CLM_FREQ              1.970e-01  2.981e-02   6.608 3.89e-11 ***
## REVOKED               8.754e-01  9.501e-02   9.214  < 2e-16 ***
## MVR_PTS               1.159e-01  1.419e-02   8.168 3.14e-16 ***
## CAR_AGE              -4.814e-03  7.615e-03  -0.632 0.527275
## URBANICITY            2.386e+00  1.131e-01  21.091  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8816.6  on 7634  degrees of freedom
## Residual deviance: 6767.4  on 7598  degrees of freedom
## AIC: 6841.4
##
## Number of Fisher Scoring iterations: 5
```

Right away, here is our base model for a logistic regression : The AIC and residual deviance *looks* high but we'll need to compare it to others. To find out for sure…I can tell by a quick glance at our p-values and significance codes that this is by far the best model. Too many features that only confound or overfit the model.

Before we throw this model away and only use it as a baseline; lets give it to our step function, to perform backward/forward stepwise.

```
## Start:  AIC=6841.41
## TARGET_FLAG ~ 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
##
##              Df Deviance    AIC
## - AGE         1   6767.6 6839.6
## - CAR_AGE     1   6767.8 6839.8
## - SEX         1   6768.6 6840.6
## - RED_CAR     1   6768.6 6840.6
```

```
## - HOMEKIDS    1   6768.7 6840.7
## - YOJ         1   6769.1 6841.1
## <none>            6767.4 6841.4
## - PARENT1     1   6776.8 6848.8
## - INCOME      1   6778.7 6850.7
## - HOME_VAL    1   6779.2 6851.2
## - OLDCLAIM    1   6780.0 6852.0
## - BLUEBOOK    1   6782.6 6854.6
## - EDUCATION   4   6788.6 6854.6
## - MSTATUS     1   6797.4 6869.4
## - KIDSDRIV    1   6801.0 6873.0
## - CLM_FREQ    1   6810.5 6882.5
## - JOB         7   6828.8 6888.8
## - TIF         1   6820.4 6892.4
## - TRAVTIME    1   6824.0 6896.0
## - MVR_PTS     1   6834.8 6906.8
## - CAR_USE     1   6837.4 6909.4
## - CAR_TYPE    5   6857.5 6921.5
## - REVOKED     1   6850.9 6922.9
## - URBANICITY  1   7406.3 7478.3
##
## Step:  AIC=6839.61
## TARGET_FLAG ~ KIDSDRIV + 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
##
##              Df Deviance    AIC
## - CAR_AGE     1   6768.0 6838.0
## - SEX         1   6768.7 6838.7
## - RED_CAR     1   6768.8 6838.8
## - HOMEKIDS    1   6769.5 6839.5
## <none>            6767.6 6839.6
## - YOJ         1   6769.6 6839.6
## + AGE         1   6767.4 6841.4
## - PARENT1     1   6777.3 6847.3
## - INCOME      1   6778.8 6848.8
## - HOME_VAL    1   6779.7 6849.7
## - OLDCLAIM    1   6780.2 6850.2
## - EDUCATION   4   6788.8 6852.8
## - BLUEBOOK    1   6783.5 6853.5
## - MSTATUS     1   6797.6 6867.6
## - KIDSDRIV    1   6801.4 6871.4
## - CLM_FREQ    1   6810.6 6880.6
## - JOB         7   6829.7 6887.7
## - TIF         1   6820.6 6890.6
## - TRAVTIME    1   6824.1 6894.1
## - MVR_PTS     1   6835.3 6905.3
## - CAR_USE     1   6837.5 6907.5
## - CAR_TYPE    5   6857.5 6919.5
## - REVOKED     1   6851.1 6921.1
## - URBANICITY  1   7407.8 7477.8
##
## Step:  AIC=6838.03
```

```
## TARGET_FLAG ~ KIDSDRIV + 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 + URBANICITY
##
##              Df Deviance    AIC
## - SEX         1   6769.2 6837.2
## - RED_CAR     1   6769.2 6837.2
## - HOMEKIDS    1   6770.0 6838.0
## - YOJ         1   6770.0 6838.0
## <none>            6768.0 6838.0
## + CAR_AGE     1   6767.6 6839.6
## + AGE         1   6767.8 6839.8
## - PARENT1     1   6777.7 6845.7
## - INCOME      1   6779.5 6847.5
## - HOME_VAL    1   6779.9 6847.9
## - OLDCLAIM    1   6780.6 6848.6
## - BLUEBOOK    1   6783.9 6851.9
## - EDUCATION   4   6793.9 6855.9
## - MSTATUS     1   6798.2 6866.2
## - KIDSDRIV    1   6801.8 6869.8
## - CLM_FREQ    1   6810.9 6878.9
## - JOB         7   6830.1 6886.1
## - TIF         1   6821.1 6889.1
## - TRAVTIME    1   6824.5 6892.5
## - MVR_PTS     1   6835.7 6903.7
## - CAR_USE     1   6838.0 6906.0
## - CAR_TYPE    5   6858.1 6918.1
## - REVOKED     1   6851.5 6919.5
## - URBANICITY  1   7408.1 7476.1
##
## Step:  AIC=6837.15
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + RED_CAR + OLDCLAIM + CLM_FREQ +
##     REVOKED + MVR_PTS + URBANICITY
##
##              Df Deviance    AIC
## - RED_CAR     1   6769.6 6835.6
## - HOMEKIDS    1   6770.9 6836.9
## - YOJ         1   6771.1 6837.1
## <none>            6769.2 6837.2
## + SEX         1   6768.0 6838.0
## + CAR_AGE     1   6768.7 6838.7
## + AGE         1   6769.0 6839.0
## - PARENT1     1   6778.8 6844.8
## - INCOME      1   6780.8 6846.8
## - HOME_VAL    1   6780.9 6846.9
## - OLDCLAIM    1   6781.8 6847.8
## - EDUCATION   4   6795.0 6855.0
## - BLUEBOOK    1   6790.7 6856.7
## - MSTATUS     1   6799.4 6865.4
## - KIDSDRIV    1   6803.0 6869.0
## - CLM_FREQ    1   6812.1 6878.1
```

```
## - JOB          7   6831.0 6885.0
## - TIF          1   6822.1 6888.1
## - TRAVTIME     1   6825.8 6891.8
## - MVR_PTS      1   6836.6 6902.6
## - CAR_USE      1   6839.2 6905.2
## - REVOKED      1   6853.0 6919.0
## - CAR_TYPE     5   6864.4 6922.4
## - URBANICITY   1   7409.4 7475.4
##
## Step:  AIC=6835.62
## TARGET_FLAG ~ KIDSDRIV + HOMEKIDS + YOJ + INCOME + PARENT1 +
##     HOME_VAL + MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE +
##     BLUEBOOK + TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED +
##     MVR_PTS + URBANICITY
##
##              Df Deviance    AIC
## - HOMEKIDS    1   6771.4 6835.4
## - YOJ         1   6771.6 6835.6
## <none>            6769.6 6835.6
## + RED_CAR     1   6769.2 6837.2
## + CAR_AGE     1   6769.2 6837.2
## + SEX         1   6769.2 6837.2
## + AGE         1   6769.5 6837.5
## - PARENT1     1   6779.4 6843.4
## - INCOME      1   6781.2 6845.2
## - HOME_VAL    1   6781.3 6845.3
## - OLDCLAIM    1   6782.3 6846.3
## - EDUCATION   4   6795.6 6853.6
## - BLUEBOOK    1   6790.9 6854.9
## - MSTATUS     1   6799.7 6863.7
## - KIDSDRIV    1   6803.7 6867.7
## - CLM_FREQ    1   6812.4 6876.4
## - JOB         7   6831.9 6883.9
## - TIF         1   6822.5 6886.5
## - TRAVTIME    1   6826.2 6890.2
## - MVR_PTS     1   6837.1 6901.1
## - CAR_USE     1   6839.7 6903.7
## - REVOKED     1   6853.4 6917.4
## - CAR_TYPE    5   6879.3 6935.3
## - URBANICITY  1   7409.7 7473.7
##
## Step:  AIC=6835.41
## TARGET_FLAG ~ KIDSDRIV + YOJ + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY
##
##              Df Deviance    AIC
## - YOJ         1   6772.9 6834.9
## <none>            6771.4 6835.4
## + HOMEKIDS    1   6769.6 6835.6
## + AGE         1   6770.7 6836.7
## + RED_CAR     1   6770.9 6836.9
## + CAR_AGE     1   6771.0 6837.0
```

```
## + SEX         1   6771.1 6837.1
## - INCOME      1   6782.6 6844.6
## - HOME_VAL    1   6783.8 6845.8
## - OLDCLAIM    1   6784.1 6846.1
## - PARENT1     1   6790.7 6852.7
## - EDUCATION   4   6797.7 6853.7
## - BLUEBOOK    1   6793.1 6855.1
## - MSTATUS     1   6799.8 6861.8
## - CLM_FREQ    1   6814.3 6876.3
## - KIDSDRIV    1   6818.9 6880.9
## - JOB         7   6834.4 6884.4
## - TIF         1   6823.9 6885.9
## - TRAVTIME    1   6827.5 6889.5
## - MVR_PTS     1   6839.4 6901.4
## - CAR_USE     1   6841.5 6903.5
## - REVOKED     1   6855.8 6917.8
## - CAR_TYPE    5   6881.5 6935.5
## - URBANICITY  1   7411.4 7473.4
##
## Step:  AIC=6834.87
## TARGET_FLAG ~ KIDSDRIV + INCOME + PARENT1 + HOME_VAL + MSTATUS +
##     EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK + TIF + CAR_TYPE +
##     OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS + URBANICITY
##
##              Df Deviance    AIC
## <none>            6772.9 6834.9
## + YOJ         1   6771.4 6835.4
## + HOMEKIDS    1   6771.6 6835.6
## + AGE         1   6771.9 6835.9
## + RED_CAR     1   6772.4 6836.4
## + CAR_AGE     1   6772.4 6836.4
## + SEX         1   6772.6 6836.6
## - INCOME      1   6785.2 6845.2
## - HOME_VAL    1   6785.3 6845.3
## - OLDCLAIM    1   6785.8 6845.8
## - PARENT1     1   6791.7 6851.7
## - EDUCATION   4   6799.0 6853.0
## - BLUEBOOK    1   6794.9 6854.9
## - MSTATUS     1   6803.6 6863.6
## - CLM_FREQ    1   6815.9 6875.9
## - KIDSDRIV    1   6819.9 6879.9
## - JOB         7   6835.2 6883.2
## - TIF         1   6825.5 6885.5
## - TRAVTIME    1   6828.7 6888.7
## - MVR_PTS     1   6841.6 6901.6
## - CAR_USE     1   6843.6 6903.6
## - REVOKED     1   6857.4 6917.4
## - CAR_TYPE    5   6883.2 6935.2
## - URBANICITY  1   7412.1 7472.1

##
## Call:
## glm(formula = TARGET_FLAG ~ KIDSDRIV + INCOME + PARENT1 + HOME_VAL +
##     MSTATUS + EDUCATION + JOB + TRAVTIME + CAR_USE + BLUEBOOK +
```

```
##     TIF + CAR_TYPE + OLDCLAIM + CLM_FREQ + REVOKED + MVR_PTS +
##     URBANICITY, family = binomial, data = m1d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5654  -0.7054  -0.3863   0.6252   3.1406
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -3.390e+00  2.001e-01 -16.943  < 2e-16 ***
## KIDSDRIV              3.900e-01  5.662e-02   6.888 5.68e-12 ***
## INCOME               -4.355e-06  1.251e-06  -3.480 0.000501 ***
## PARENT1               4.228e-01  9.741e-02   4.340 1.42e-05 ***
## HOME_VAL             -1.348e-06  3.821e-07  -3.527 0.000421 ***
## MSTATUS              -4.754e-01  8.523e-02  -5.578 2.44e-08 ***
## EDUCATIONBachelors   -3.705e-01  1.097e-01  -3.377 0.000733 ***
## EDUCATIONHigh School  1.605e-02  9.493e-02   0.169 0.865734
## EDUCATIONMasters     -3.381e-01  1.669e-01  -2.026 0.042814 *
## EDUCATIONPhD          4.042e-02  2.192e-01   0.184 0.853725
## JOBClerical           9.202e-02  1.074e-01   0.857 0.391556
## JOBDoctor            -9.017e-01  3.016e-01  -2.990 0.002792 **
## JOBHome Maker        -8.703e-02  1.469e-01  -0.592 0.553641
## JOBLawyer            -1.491e-01  1.893e-01  -0.788 0.430910
## JOBManager           -8.594e-01  1.400e-01  -6.139 8.28e-10 ***
## JOBProfessional      -1.348e-01  1.203e-01  -1.120 0.262658
## JOBStudent           -8.724e-02  1.256e-01  -0.695 0.487277
## TRAVTIME              1.452e-02  1.947e-03   7.459 8.70e-14 ***
## CAR_USE               7.802e-01  9.343e-02   8.351  < 2e-16 ***
## BLUEBOOK             -2.304e-05  4.954e-06  -4.651 3.30e-06 ***
## TIF                  -5.446e-02  7.641e-03  -7.127 1.02e-12 ***
## CAR_TYPEPanel Truck   6.397e-01  1.658e-01   3.857 0.000115 ***
## CAR_TYPEPickup        5.787e-01  1.024e-01   5.652 1.58e-08 ***
## CAR_TYPESports Car    9.854e-01  1.081e-01   9.120  < 2e-16 ***
## CAR_TYPESUV           7.286e-01  8.657e-02   8.416  < 2e-16 ***
## CAR_TYPEVan           5.987e-01  1.284e-01   4.663 3.11e-06 ***
## OLDCLAIM             -1.460e-05  4.100e-06  -3.562 0.000368 ***
## CLM_FREQ              1.965e-01  2.978e-02   6.601 4.09e-11 ***
## REVOKED               8.796e-01  9.490e-02   9.269  < 2e-16 ***
## MVR_PTS               1.169e-01  1.417e-02   8.251  < 2e-16 ***
## URBANICITY            2.385e+00  1.131e-01  21.091  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8816.6  on 7634  degrees of freedom
## Residual deviance: 6772.9  on 7604  degrees of freedom
## AIC: 6834.9
##
## Number of Fisher Scoring iterations: 5
```

Everything went up; + it dropped almost no features. We're throwing out these two models and we'll use the data I've modified.

```
##
```

```
## Call:
## glm(formula = TARGET_FLAG ~ ., family = binomial, data = m2d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2755  -0.7192  -0.3893   0.6580   3.2418
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.582e+00  6.425e-01  -2.462 0.013817 *
## KIDSDRIV              5.599e-01  9.791e-02   5.719 1.07e-08 ***
## AGE                  -4.851e-05  4.246e-03  -0.011 0.990884
## HOMEKIDS              1.846e-01  8.657e-02   2.132 0.032972 *
## YOJ                  -8.967e-02  5.805e-02  -1.545 0.122448
## INCOME               -4.609e-06  1.190e-06  -3.872 0.000108 ***
## PARENT1               2.646e-01  1.198e-01   2.209 0.027191 *
## HOME_VAL             -2.391e-01  5.734e-02  -4.171 3.04e-05 ***
## MSTATUS              -5.038e-01  9.045e-02  -5.570 2.55e-08 ***
## SEX                   1.545e-01  1.130e-01   1.368 0.171467
## EDUCATIONBachelors   -3.167e-01  1.167e-01  -2.713 0.006666 **
## EDUCATIONHigh School  2.790e-02  9.494e-02   0.294 0.768835
## EDUCATIONMasters     -2.644e-01  1.754e-01  -1.507 0.131720
## EDUCATIONPhD          7.313e-02  2.249e-01   0.325 0.745073
## JOBClerical           6.191e-02  1.071e-01   0.578 0.563188
## JOBDoctor            -8.590e-01  2.994e-01  -2.869 0.004117 **
## JOBHome Maker        -1.108e-01  1.525e-01  -0.726 0.467722
## JOBLawyer            -1.309e-01  1.896e-01  -0.690 0.489890
## JOBManager           -8.498e-01  1.403e-01  -6.056 1.39e-09 ***
## JOBProfessional      -1.137e-01  1.202e-01  -0.946 0.344009
## JOBStudent           -1.821e-01  1.298e-01  -1.403 0.160619
## TRAVTIME              4.053e-01  5.289e-02   7.664 1.80e-14 ***
## CAR_USE               7.919e-01  9.333e-02   8.485  < 2e-16 ***
## BLUEBOOK             -2.811e-01  5.959e-02  -4.717 2.39e-06 ***
## TIF                  -4.152e-01  6.711e-02  -6.187 6.13e-10 ***
## CAR_TYPEPanel Truck   5.211e-01  1.636e-01   3.186 0.001442 **
## CAR_TYPEPickup        5.849e-01  1.017e-01   5.750 8.93e-09 ***
## CAR_TYPESports Car    1.022e+00  1.291e-01   7.912 2.53e-15 ***
## CAR_TYPESUV           7.968e-01  1.090e-01   7.310 2.68e-13 ***
## CAR_TYPEVan           5.616e-01  1.313e-01   4.277 1.89e-05 ***
## RED_CAR              -9.661e-02  9.203e-02  -1.050 0.293854
## OLDCLAIM              3.409e-02  6.133e-02   0.556 0.578365
## CLM_FREQ              2.576e-01  7.379e-02   3.491 0.000481 ***
## REVOKED               7.151e-01  8.446e-02   8.467  < 2e-16 ***
## MVR_PTS               2.258e-01  3.776e-02   5.981 2.22e-09 ***
## CAR_AGE              -3.841e-02  4.255e-02  -0.903 0.366634
## URBANICITY            2.357e+00  1.121e-01  21.033  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8816.6  on 7634  degrees of freedom
## Residual deviance: 6800.5  on 7598  degrees of freedom
## AIC: 6874.5
```

```
##
## Number of Fisher Scoring iterations: 5
```

Dissapointing we didn't have an enormous decrease in AIC or deviance, but it is better than the base model. Lets try stepwise.

```
##
## Call:
## glm(formula = TARGET_FLAG ~ INCOME + HOME_VAL + MSTATUS + TRAVTIME +
##     BLUEBOOK + CAR_USE + URBANICITY + TIF + CLM_FREQ + REVOKED,
##     family = binomial, data = m2d)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3083  -0.7567  -0.4402   0.7495   2.9005
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.071e-01  4.897e-01   0.831    0.406
## INCOME      -9.206e-06  9.037e-07 -10.188  < 2e-16 ***
## HOME_VAL    -2.423e-01  5.321e-02  -4.553 5.28e-06 ***
## MSTATUS     -5.323e-01  7.132e-02  -7.463 8.45e-14 ***
## TRAVTIME     3.795e-01  5.110e-02   7.427 1.11e-13 ***
## BLUEBOOK    -3.943e-01  4.900e-02  -8.046 8.55e-16 ***
## CAR_USE      9.080e-01  6.134e-02  14.802  < 2e-16 ***
## URBANICITY   2.161e+00  1.092e-01  19.785  < 2e-16 ***
## TIF         -4.055e-01  6.490e-02  -6.247 4.18e-10 ***
## CLM_FREQ     4.075e-01  3.996e-02  10.199  < 2e-16 ***
## REVOKED      7.754e-01  8.043e-02   9.640  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8816.6  on 7634  degrees of freedom
## Residual deviance: 7180.4  on 7624  degrees of freedom
## AIC: 7202.4
##
## Number of Fisher Scoring iterations: 5
```

With AIC and deviance going up by neglible amounts, while dropping 2/3rds of the predictors. Looks like a decent model.

Lets move on to linear models

```
##
## Call:
## lm(formula = TARGET_AMT ~ ., data = l1d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -6201   -440    -50    228 101040
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -5.402e+02  4.022e+02  -1.343  0.17932
```

```
## TARGET_FLAG           5.636e+03  1.152e+02  48.927  < 2e-16 ***
## KIDSDRIV             -7.380e+01  9.874e+01  -0.747  0.45483
## AGE                   4.410e+00  6.219e+00   0.709  0.47830
## HOMEKIDS              1.403e+00  5.703e+01   0.025  0.98038
## YOJ                   1.270e+01  1.282e+01   0.991  0.32169
## INCOME                1.899e-04  1.762e-03   0.108  0.91416
## PARENT1               9.910e+01  1.778e+02   0.557  0.57737
## HOME_VAL             -5.068e-04  5.638e-04  -0.899  0.36874
## MSTATUS              -6.592e+01  1.323e+02  -0.498  0.61821
## SEX                   2.594e+02  1.621e+02   1.600  0.10957
## EDUCATIONBachelors    2.476e+01  1.757e+02   0.141  0.88796
## EDUCATIONHigh School -1.387e+02  1.471e+02  -0.943  0.34583
## EDUCATIONMasters      1.316e+02  2.610e+02   0.504  0.61414
## EDUCATIONPhD          3.770e+02  3.267e+02   1.154  0.24858
## JOBClerical          -4.439e+01  1.657e+02  -0.268  0.78877
## JOBDoctor            -4.438e+02  3.908e+02  -1.135  0.25621
## JOBHome Maker        -6.056e+01  2.320e+02  -0.261  0.79411
## JOBLawyer            -1.964e+00  2.679e+02  -0.007  0.99415
## JOBManager           -2.120e+02  2.018e+02  -1.050  0.29353
## JOBProfessional       1.135e+02  1.830e+02   0.620  0.53526
## JOBStudent           -1.686e+02  2.018e+02  -0.835  0.40352
## TRAVTIME             -1.674e+00  2.843e+00  -0.589  0.55605
## CAR_USE               1.137e+02  1.441e+02   0.789  0.43018
## BLUEBOOK              2.461e-02  7.626e-03   3.227  0.00125 **
## TIF                  -7.072e+00  1.081e+01  -0.654  0.51312
## CAR_TYPEPanel Truck  -9.307e+01  2.592e+02  -0.359  0.71960
## CAR_TYPEPickup       -6.953e+01  1.486e+02  -0.468  0.63990
## CAR_TYPESports Car    2.239e+02  1.882e+02   1.189  0.23432
## CAR_TYPESUV           1.696e+02  1.554e+02   1.092  0.27500
## CAR_TYPEVan           2.152e+02  1.895e+02   1.136  0.25614
## RED_CAR               2.316e+01  1.350e+02   0.172  0.86374
## OLDCLAIM              2.562e-03  6.625e-03   0.387  0.69898
## CLM_FREQ             -7.881e+01  4.914e+01  -1.604  0.10884
## REVOKED              -3.456e+02  1.541e+02  -2.243  0.02494 *
## MVR_PTS               6.032e+01  2.308e+01   2.614  0.00898 **
## CAR_AGE              -2.108e+01  1.094e+01  -1.928  0.05395 .
## URBANICITY           -8.691e+00  1.253e+02  -0.069  0.94469
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3879 on 7597 degrees of freedom
## Multiple R-squared:  0.2948, Adjusted R-squared:  0.2914
## F-statistic: 85.84 on 37 and 7597 DF,  p-value: < 2.2e-16
```

R2 coefficient sitting at approx. .3 with a great p-value and OK f-statistic. Using everything we already have by default + tidying it looks like an acceptable model.

Next up; were going to only use the significant features and see what happens.

```
##
## Call:
## lm(formula = TARGET_AMT ~ TARGET_FLAG + BLUEBOOK + MVR_PTS, data = l1d)
##
## Residuals:
##    Min     1Q Median     3Q    Max
```

```
##  -6036   -309    -25    164 101633
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.628e+02  1.064e+02  -4.352 1.37e-05 ***
## TARGET_FLAG  5.619e+03  1.039e+02  54.100  < 2e-16 ***
## BLUEBOOK     2.514e-02  5.572e-03   4.512 6.51e-06 ***
## MVR_PTS      4.883e+01  2.127e+01   2.296   0.0217 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3879 on 7631 degrees of freedom
## Multiple R-squared:  0.2916, Adjusted R-squared:  0.2913
## F-statistic:  1047 on 3 and 7631 DF,  p-value: < 2.2e-16
```

Same R2, same great p-value. F-statistic sky rocketed which is OK in this situation since with only 3 features I don't think we're overfitting. I'm sad to see MVR_PTS become insignifcant. Lets see if we can make the TARGET_AMOUNT a little more normal by logging it and taking away MVR_PTS.

```
##
## Call:
## lm(formula = log(TARGET_AMT + 1) ~ TARGET_FLAG + BLUEBOOK, data = l1d)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8086 -0.0172  0.0032  0.0199  3.3089
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.727e-02  1.092e-02  -3.413 0.000647 ***
## TARGET_FLAG  8.271e+00  1.089e-02 759.387  < 2e-16 ***
## BLUEBOOK     2.375e-06  5.989e-07   3.966 7.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4171 on 7632 degrees of freedom
## Multiple R-squared:  0.9871, Adjusted R-squared:  0.9871
## F-statistic: 2.916e+05 on 2 and 7632 DF,  p-value: < 2.2e-16
```

Interesting R2 at .99 with a crazy high F-statistic. I'd like to say I'm overfitting a model but perhaps claims are settled by their bluebooked value. Just for fun lets see what my modified dataset could have done in one more model.

```
##
## Call:
## lm(formula = TARGET_AMT ~ TARGET_FLAG + SEX + REVOKED + BLUEBOOK,
##     data = train2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -5703   -314    -87    156 101620
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3659.56     681.66  -5.369 8.17e-08 ***
## TARGET_FLAG  5722.18     102.43  55.862  < 2e-16 ***
```

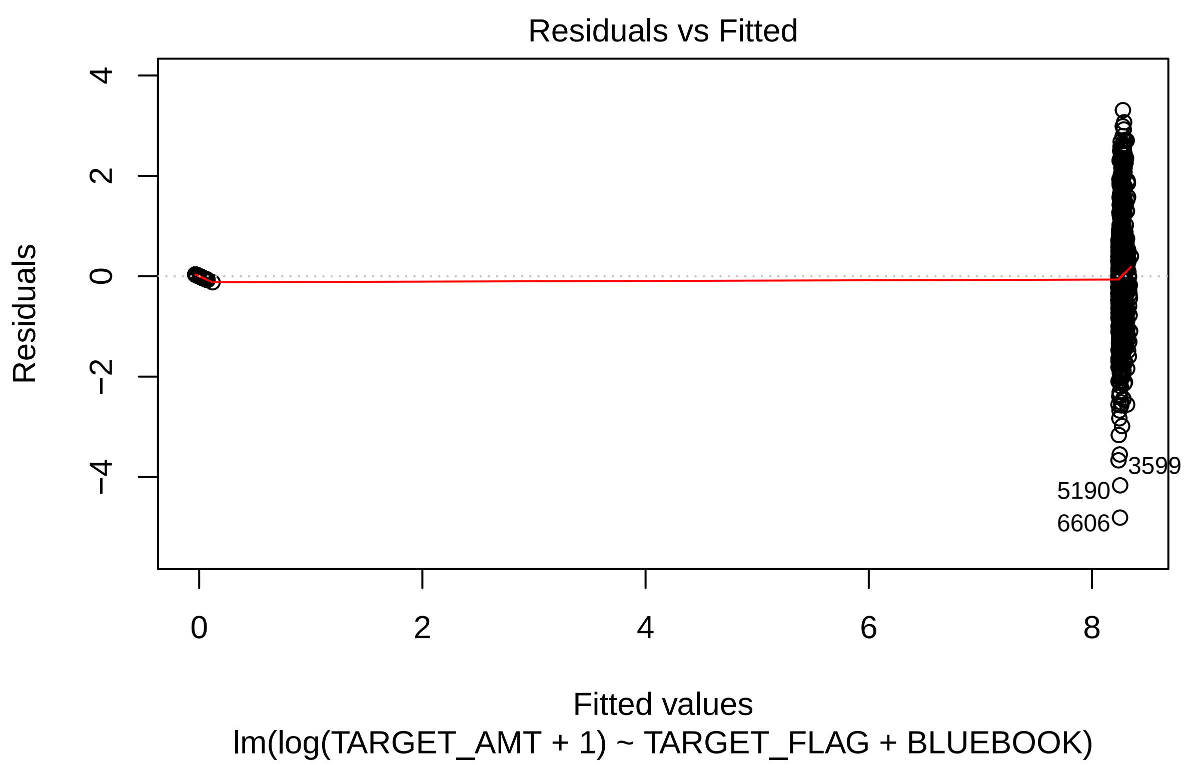
```
## SEX            185.86      89.42   2.078   0.0377 *
## REVOKED       -304.15     136.72  -2.225   0.0261 *
## BLUEBOOK       379.21      71.49   5.304 1.16e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3876 on 7630 degrees of freedom
## Multiple R-squared:  0.2928, Adjusted R-squared:  0.2925
## F-statistic: 789.9 on 4 and 7630 DF,  p-value: < 2.2e-16
```

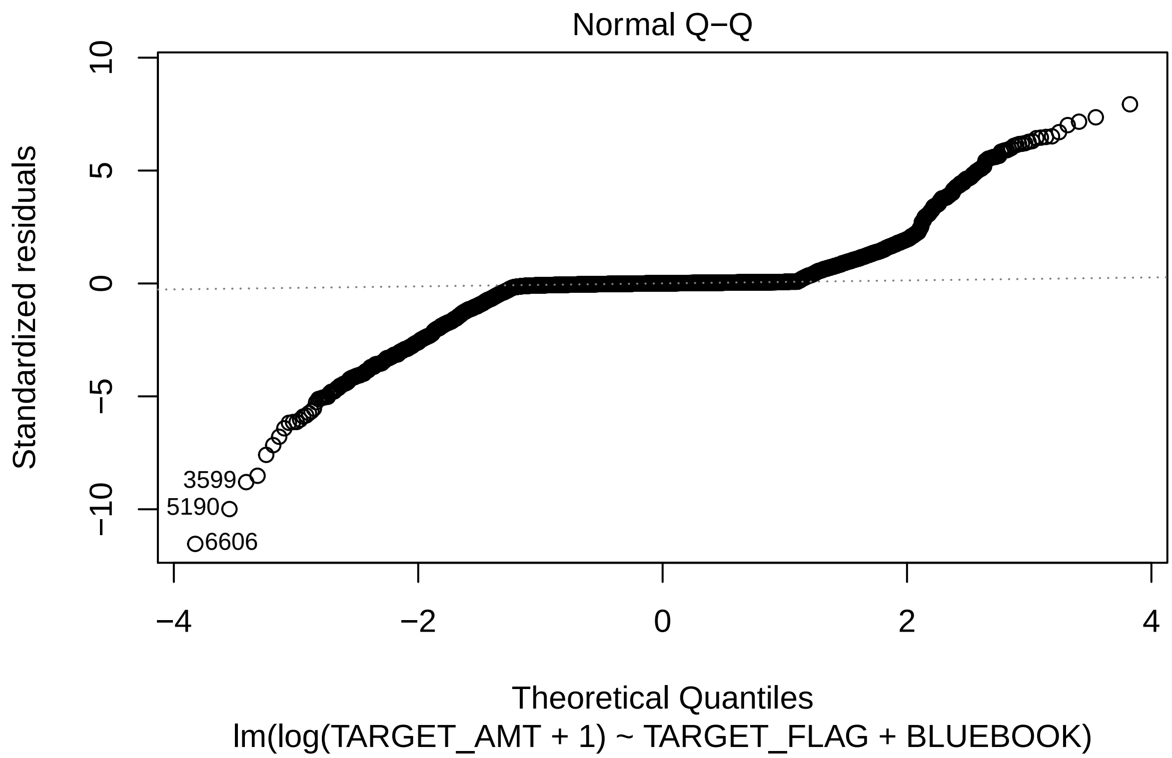
So by standard linear model metrics alone; it would seem my transformations and imputations were beneficial to a base model/dataset. The advantage of my modified dataset is that most of the calculations are factor based which helps counter-overfitting, nice!

4. SELECT MODELS

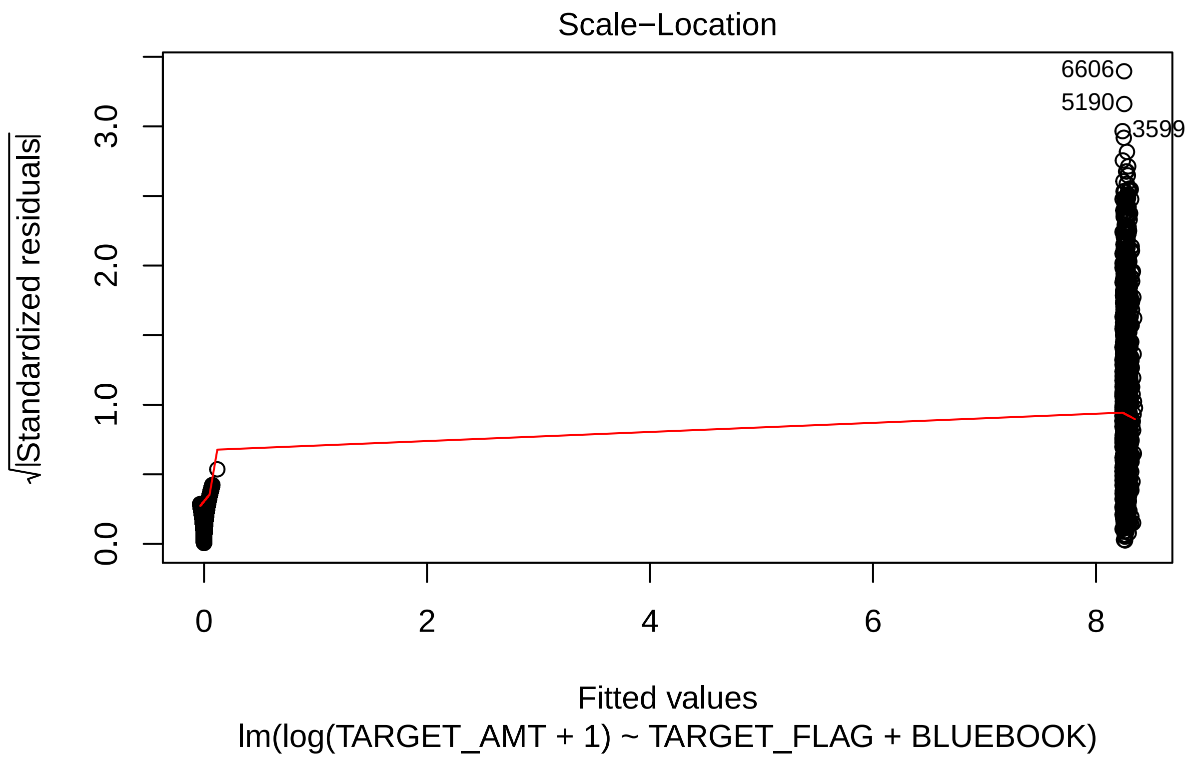
I choose logistic model 1 because of my failure to further increase the strength of the models significantly through data transformation. If my transformations don't help the model fit more than they hinder any chance of replication, they are better off left alone.

I also choose linear model 3. Lets evaluate them.

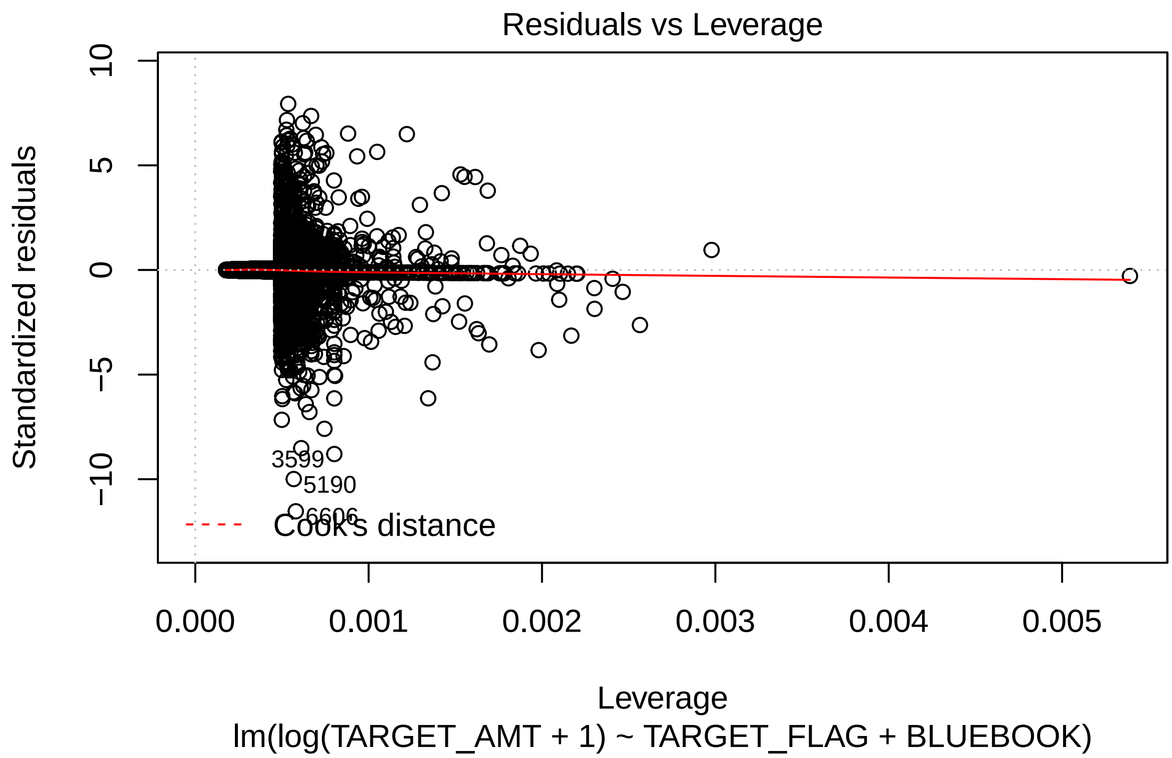




Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log(TARGET_AMT + 1) ~ TARGET_FLAG + BLUEBOOK)
3599
5190
6606



Scale−Location
√|Standardized residuals|
0.0
1.0
2.0
3.0
0
2
4
6
8
6606
5190
3599
Fitted values
lm(log(TARGET_AMT + 1) ~ TARGET_FLAG + BLUEBOOK)



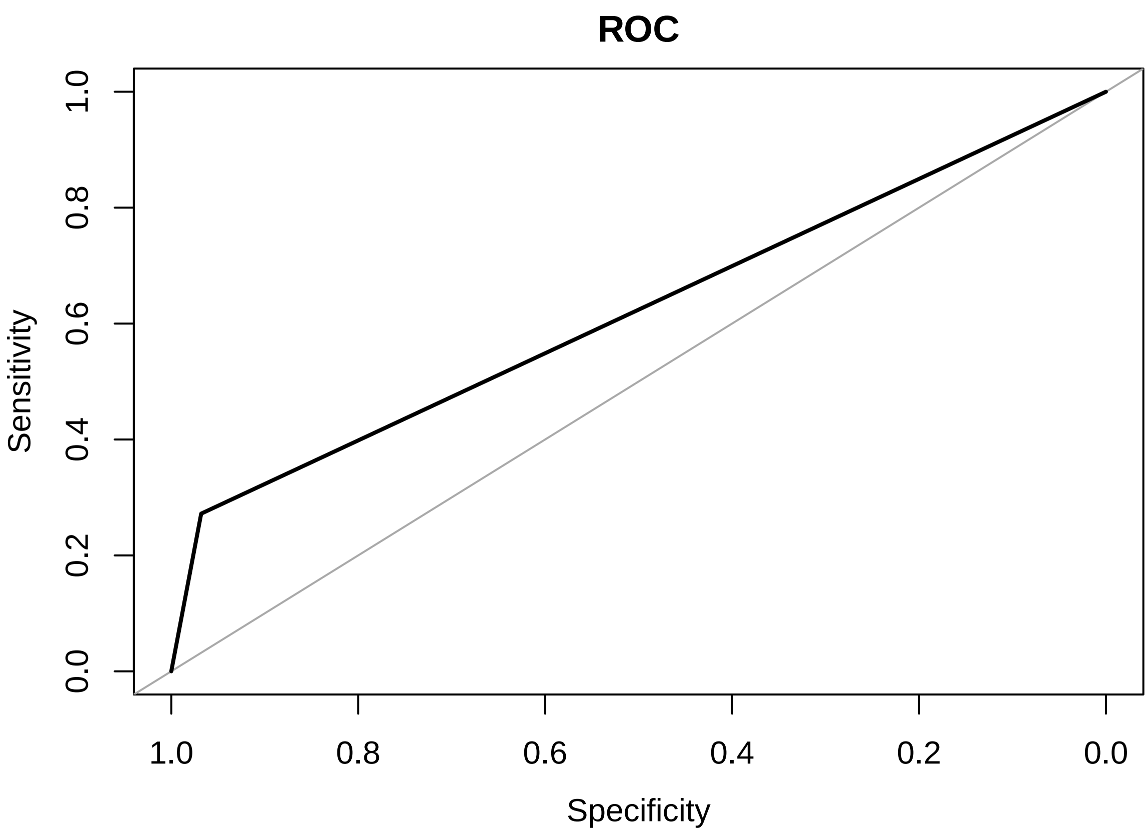
Looking at the normal Q-Q plot of our linear model. I see the model does not rely on normalacy whatsoever. Either the bluebook has a direct, almost 1 to 1 correlation on claim amounts, or there is a flaw in my methodology. In which I would go for linear model 4 with a .3 R2 and a great F-statistic; using just a little more variables but keeping it parsimonious.

Lets look at the binary model.

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 2177  586
##          1   72  219
##
##                Accuracy : 0.7845
##                  95% CI : (0.7695, 0.799)
##     No Information Rate : 0.7364
##     P-Value [Acc > NIR] : 4.034e-10
##
##                   Kappa : 0.3019
##  Mcnemar's Test P-Value : < 2.2e-16
##
##             Sensitivity : 0.27205
##             Specificity : 0.96799
##          Pos Pred Value : 0.75258
##          Neg Pred Value : 0.78791
##              Prevalence : 0.26359
```

```
##          Detection Rate : 0.07171
##    Detection Prevalence : 0.09528
##       Balanced Accuracy : 0.62002
##
##        'Positive' Class : 1
##
```

With an accuracy of .79 and high specificity. This model is acceptable, it is a shame I could not significantly improve it from the base I am using now.




```
## Area under the curve: 0.62
```

Code Appendix

```
knitr::opts_chunk$set(echo = FALSE)
#https://stackoverflow.com/questions/9341635/check-for-installed-packages-before-running-install-packag
requiredPackages = c('knitr','prettydoc','kableExtra','ggplot2','tidyr','plyr','dplyr','psych','corrplo
for(p in requiredPackages){
  if(!require(p,character.only = TRUE)) install.packages(p)
  library(p,character.only = TRUE)
}
#Load Data
train = read.csv('insurance_training_data.csv',na.strings=c(""," ","NA"))
eval = read.csv('insurance_evaluation_data.csv',na.strings=c(""," ","NA"))
toBinary = function(aVector,one){
  return(ifelse(aVector==one,1,0))
```

```
}
toClean = function(aVector,isNumeric){
  if(isNumeric==TRUE){
    return(as.numeric(gsub('\\$|,','',aVector)))
  }
  return(sub('z_|Z_','',aVector))
}
toDataset = function(d){
  d$INDEX = NULL
  d$INCOME = toClean(d$INCOME,TRUE)
  d$PARENT1 = toBinary(d$PARENT1,'Yes')
  d$MSTATUS = toBinary(d$MSTATUS,'Yes')
  d$HOME_VAL = toClean(d$HOME_VAL,TRUE)
  d$SEX = toBinary(d$SEX,'M')
  d$EDUCATION = toClean(d$EDUCATION,FALSE)
  d$JOB = toClean(d$JOB,FALSE)
  d$CAR_USE = toBinary(d$CAR_USE,'Commercial')
  d$BLUEBOOK = toClean(d$BLUEBOOK,TRUE)
  d$CAR_TYPE = toClean(d$CAR_TYPE,FALSE)
  d$RED_CAR = toBinary(d$RED_CAR,'yes')
  d$OLDCLAIM = toClean(d$OLDCLAIM,TRUE)
  d$REVOKED = toBinary(d$REVOKED,'Yes')
  d$URBANICITY = toBinary(d$URBANICITY,'Highly Urban/ Urban')
  return(d)
}
train = toDataset(train)
eval = toDataset(eval)
#Using the mice package to identify the missing variables again
missingValuePlot = aggr(train, col=c('orange','yellow'),
                    numbers=TRUE, sortVars=TRUE, only.miss=TRUE, combined=TRUE,
                    labels=names(train), cex.axis=.4,
                    gap=3, ylab=c("Missing data","Pattern"))
#https://stackoverflow.com/questions/14604439/plot-multiple-boxplot-in-one-graph/14606549
theNumerics = c('KIDSDRIV','HOMEKIDS','YOJ','AGE','INCOME','HOME_VAL','TRAVTIME','BLUEBOOK','TIF','OLDC
notNumerics = c('PARENT1','MSTATUS','SEX','EDUCATION','JOB','CAR_USE','CAR_TYPE','RED_CAR','REVOKED','U
nonNum = train[,notNumerics]
theNums = train[,theNumerics]
train2 = train[,theNumerics]
train2$TARGET_FLAG = train$TARGET_FLAG
train2$TARGET_FLAG = ifelse(train$TARGET_FLAG==1,'Crash','No Crash')
train.m = melt(train2,id.var='TARGET_FLAG')
require(ggplot2)
p = ggplot(data = train.m, aes(x=variable, y=value))
p = p + geom_boxplot(aes(fill = TARGET_FLAG))
# if you want color for points replace group with colour=Label
p = p + facet_wrap( ~ variable, scales="free")
p = p + xlab("x-axis") + ylab("y-axis") + ggtitle("Variance in target ~ numeric feature")
p = p + guides(fill=guide_legend(title="Legend_Title"))
p
library(Hmisc)
train2$TARGET_FLAG = NULL
hist(train2[1:6],na.big = FALSE)
hist(train2[7:13],na.big=FALSE)
```

```
#Creating a correlation matrix to address multi-colinearity issues
train3 = train2
train3$TARGET_FLAG = train$TARGET_FLAG
train3$TARGET_AMG = train$TARGET_AMT
correlationMatrix = cor(train3, use='complete.obs')
corrplot(correlationMatrix, method="pie")
mosaicplot(table(train$JOB,train$EDUCATION),col = hcl(c(110, 50)),main= 'Job vs Education')
mosaicplot(table(train$AGE,train$YOJ),col = hcl(c(190, 10)),main= 'Age vs Years on Job')
imputedData = mice(train, m=2, maxit = 5, method = 'pmm', seed = 15)
train=mice::complete(imputedData,2)
train = train[complete.cases(train), ]
missingValuePlot = aggr(train, col=c('orange','yellow'),
                    numbers=TRUE, sortVars=TRUE, only.miss=FALSE, combined=TRUE,
                    labels=names(train), cex.axis=.4,
                    gap=3, ylab=c("Missing data","Pattern"))
 #HOMEKIDS, HOME_VAL,TIF,MVR_PTS, CLM_FREQ,OLDCLAIM,CAR_AGE, YOJ
train2 = train
train2$KIDSDRIV = ifelse(train$KIDSDRIV==0,0,1)
train2$HOMEKIDS = ifelse(train$HOMEKIDS==0,0,ifelse(train$HOMEKIDS>3,2,1))
train2$HOME_VAL = ifelse(train$HOME_VAL==0,0,ifelse(train$HOME_VAL<quantile(train$HOME_VAL,.75),1,2))
train2$TIF = ifelse(train$TIF==0,0,ifelse(train$TIF<quantile(train$TIF,.75),1,2))
train2$MVR_PTS = ifelse(train$MVR_PTS==0,0,ifelse(train$MVR_PTS<quantile(train$MVR_PTS,.75),1,2))
train2$CLM_FREQ = ifelse(train$CLM_FREQ==0,0,ifelse(train$CLM_FREQ<3,1,2))
train2$OLDCLAIM = ifelse(train$OLDCLAIM==0,0,ifelse(train$OLDCLAIM<quantile(train$OLDCLAIM,.75),1,2))
train2$CAR_AGE = ifelse(train$CAR_AGE==1,1,ifelse(train$CAR_AGE<quantile(train$CAR_AGE,.5),2,3))
train2$YOJ = ifelse(train$YOJ==0,0,ifelse(train$YOJ<quantile(train$YOJ,.75),1,2))
train2$BLUEBOOK = log(train$BLUEBOOK)
train2$TRAVTIME = log(train$TRAVTIME)

library(Hmisc)
newHistData = train2[,theNumerics]
hist(newHistData[1:6],na.big = FALSE)
hist(newHistData[7:13],na.big=FALSE)
#Model 1 : Original dataset
m1d = train
m1d$TARGET_AMT = NULL
m1 = glm(data=m1d,TARGET_FLAG~.,family=binomial)
summary(m1)
#Model 2 : Original dataset + Stepwise model fitting

m2 = step(m1,direction = 'both')
summary(m2)
#Model 3 : modified dataset
m2d = train2
m2d$TARGET_AMT = NULL
m3 = glm(data=m2d,TARGET_FLAG~.,family=binomial)
summary(m3)
#Model 4 : modified dataset Stepwise
 #HOMEKIDS, HOME_VAL,TIF,MVR_PTS, CLM_FREQ,OLDCLAIM,CAR_AGE, YOJ
m3 = glm(data=m2d,TARGET_FLAG~  INCOME+HOME_VAL+MSTATUS+TRAVTIME+BLUEBOOK+CAR_USE+URBANICITY+TIF+CLM_FR
summary(m3)

l1d = train
```

```r
#l1d = l1d[ which(l1d$TARGET_FLAG==1),]
#l1d$TARGET_FLAG = NULL
l1 = lm(data=l1d,TARGET_AMT~.)
summary(l1)

l2 = lm(data=l1d,TARGET_AMT~TARGET_FLAG+BLUEBOOK+MVR_PTS)
summary(l2)
l3 = lm(data=l1d,log(TARGET_AMT+1)~TARGET_FLAG+BLUEBOOK)
summary(l3)
l4 = lm(data=train2,TARGET_AMT~TARGET_FLAG+SEX+REVOKED+BLUEBOOK)
summary(l4)
plot(l3)
#https://www.r-bloggers.com/evaluating-logistic-regression-models/
evalTest = m1d
evalTest$TARGET_AMT = NULL
trainE = createDataPartition(evalTest$TARGET_FLAG,p=.6,list=FALSE)
trainingData = train[ trainE, ]
testingData = train[ -trainE, ]
pred = predict(m1, newdata=testingData)
pred = ifelse(pred<.5,0,1)
theMatrix = confusionMatrix(data=pred,testingData$TARGET_FLAG,positive = '1')
theMatrix
theRock = roc(testingData$TARGET_FLAG, pred)
plot(theRock,asp=NA,main='ROC')
theRock$auc
#Load Data
finalPred1 =predict(m1, eval)
probs = finalPred1
classes = ifelse(finalPred1<.5,0,1)
eval2=eval
eval2$TARGET_FLAG = 1
cost = predict(l3, eval2)
answers = cbind(probs,classes,cost)
write.csv(answers,'MMullerPredictions.csv')
```