# Assignment 4

*Kai Lukowiak*

*2018-03-25*

**Abstract**

This paper looks into the predictive ability of certain factors into the likelyhood of a person getting into an accident and also the amount that the accident will cost.



# 1 Data Exploration

## 1.1 The Data Frames

| TARGET_FLAG | TARGET_AMT | KIDSDRIV | AGE | HOMEKIDS | YOJ | INCOME | PARENT1 | HOME_VAL | MSTATUS | SEX | EDUCATION | JOB | TRAVTIME |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.000 | 0 | 30 | 2 | 13 | 103881 | No | 287047 | Yes | M | Bachelors | z_Blue Collar | 33 |
| 1 | 2044.217 | 0 | 33 | 0 | 9 | 69641 | No | 230290 | z_No | M | Bachelors | z_Blue Collar | 33 |
| 0 | 0.000 | 0 | 46 | 0 | 12 | 88359 | No | 267679 | Yes | z_F | Masters | Lawyer | 25 |
| 0 | 0.000 | 0 | 27 | 3 | 7 | 131567 | Yes | 0 | z_No | z_F | Bachelors | Professional | 32 |
| 0 | 0.000 | 0 | 51 | 0 | 13 | 85345 | No | NA | Yes | M | Masters | NA | 37 |
| 0 | 0.000 | 0 | 35 | 1 | 10 | 16039 | No | 124191 | Yes | z_F | z_High School | Clerical | 5 |

```
## Observations: 6,121
## Variables: 25
## $ TARGET_FLAG <int> 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 1, 0, 0, 1, 1, 0, 0,...
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 0.000, 0.000, 2946.000, 0.000...
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
## $ AGE         <int> 60, 43, 35, 51, 50, 34, 54, 37, 34, 50, 53, 43, 55...
## $ HOMEKIDS    <int> 0, 0, 1, 0, 0, 1, 0, 2, 0, 0, 0, 0, 0, 0, 0, 3, 3,...
## $ YOJ         <int> 11, 11, 10, 14, NA, 12, NA, NA, 10, 7, 14, 5, 11, ...
## $ INCOME      <dbl> 67349, 91449, 16039, NA, 114986, 125301, 18755, 10...
## $ PARENT1     <fct> No, No, No, No, No, Yes, No, No, No, No, No, No, N...
## $ HOME_VAL    <dbl> 0, 257252, 124191, 306251, 243925, 0, NA, 333680, ...
## $ MSTATUS     <fct> z_No, z_No, Yes, Yes, Yes, z_No, Yes, Yes, z_No, z...
## $ SEX         <fct> M, M, z_F, M, z_F, z_F, z_F, M, z_F, M, z_F, z_F, ...
## $ EDUCATION   <fct> PhD, z_High School, z_High School, <High School, P...
## $ JOB         <fct> Professional, z_Blue Collar, Clerical, z_Blue Coll...
## $ TRAVTIME    <int> 14, 22, 5, 32, 36, 46, 33, 44, 34, 48, 15, 36, 25,...
## $ CAR_USE     <fct> Private, Commercial, Private, Private, Private, Co...
## $ BLUEBOOK    <dbl> 14230, 14940, 4010, 15440, 18000, 17430, 8780, 169...
```

```
## $ TIF         <int> 11, 1, 4, 7, 1, 1, 1, 1, 1, 7, 1, 7, 7, 6, 1, 6, 7...
## $ CAR_TYPE    <fct> Minivan, Minivan, z_SUV, Minivan, z_SUV, Sports Ca...
## $ RED_CAR     <fct> yes, yes, no, yes, no, no, no, yes, no, no, no, no...
## $ OLDCLAIM    <dbl> 4461, 0, 38690, 0, 19217, 0, 0, 2374, 0, 0, 0, 0, ...
## $ CLM_FREQ    <int> 2, 0, 2, 0, 2, 0, 0, 1, 0, 0, 0, 0, 2, 0, 0, 0, 0,...
## $ REVOKED     <fct> No, No, No, No, Yes, No, No, Yes, No, No, No, No, ...
## $ MVR_PTS     <int> 3, 0, 3, 0, 3, 0, 0, 10, 0, 1, 0, 0, 3, 3, 3, 0, 0...
## $ CAR_AGE     <int> 18, 1, 10, 6, 17, 7, 1, 7, 1, 17, 11, 1, 9, 10, 5,...
## $ URBANICITY  <fct> Highly Urban/ Urban, Highly Urban/ Urban, Highly U...
```

The trainind dataset is comprised of 26 variables, two of which are response variables, `TARGET_FLAG` and `TARGET_AMT`. These will be used to run logistic and regular regression respectivly.

The evaluation set looks similar but has `NA`s in the first two rows.

Sample of Values for the Test Set

| KIDSDRIV | AGE | HOMEKIDS | YOJ | INCOME | PARENT1 | HOME_VAL | MSTATUS |
|---|---|---|---|---|---|---|---|
| 0 | 47 | 0 | 12 | 36150 | No | 174964 | Yes |
| 0 | 57 | 0 | 10 | 130831 | No | 365792 | Yes |
| 0 | 50 | 0 | 14 | 15989 | No | 117038 | Yes |
| 0 | 47 | 0 | 14 | 2825 | No | 91520 | Yes |
| 0 | 58 | 0 | 12 | 73283 | No | 251025 | Yes |
| 0 | 39 | 3 | 14 | 51524 | Yes | 0 | z_No |

The evaluation set is a similar data frame but excludes the target variable. As such it cannot be used for cross validation.

```
## Parsed with column specification:
## cols(
##   `Variable Name` = col_character(),
##   Definition = col_character(),
##   `Theoretical Effect` = col_character()
## )
```

| Variable Name | Definition | Theoretical Effect |
|---|---|---|
| INDEX | Index | None |
| TARGET_FLAG | Identification Variable (do not use) | None |
| TARGET_AMT | Was Car in crash 1=YES 0=NO | None |
| AGE | Age of Driver | Very young people tend to be risky. Maybe very old people also. |
| BLUEBOOK | Value of Vehicle | Unknown effect on probability of collision, but probably effect the payout if there is a crash |
| CAR_AGE | Vehicle Age | Unknown effect on probability of collision, but probably effect the payout if there is a crash |
| CAR_TYPE | Type of Car | Unknown effect on probability of collision, but probably effect the payout if there is a crash |
| CAR_USE | Vehicle Use | Commercial vehicles are driven more, so might increase probability of collision |
| CLM_FREQ | # Claims (Past 5 Years) | The more claims you filed in the past, the more you are likely to file in the future |
| EDUCATION | Max Education Level | Unknown effect, but in theory more educated people tend to drive more safely |
| HOMEKIDS | # Children at Home | Unknown effect |
| HOME_VAL | Home Value | In theory, home owners tend to drive more responsibly |
| INCOME | Income | In theory, rich people tend to get into fewer crashes |
| JOB | Job Category | In theory, white collar jobs tend to be safer |
| KIDSDRIV | # Driving Children | When teenagers drive your car, you are more likely to get into crashes |
| MSTATUS | Matitial Status | In theory, married people drive more safely |
| MVR_PTS | Motor Vehicle Record Points | If you get lots of traffic tickets, you tend to get into more crashes |
| OLDCLAIM | Total Claims (Past 5 Years) | If your total payout over the past five years was high, this suggests future payouts will be high |
| PARENT1 | Single Parent | Unknown effect |
| RED_CAR | A red Car | Urban legend says that red cars (especially red sports cars) are more risky. Is that true? |
| REVOKED | License Revoked (Past 7 Years) | If your license was revoked in the past 7 years, you probably are a more risky driver. |
| Sex | Time in Force | Urban legend says that women have less crashes then men. Is that true? |
| TIF | Time in Force | People who have been customers for a long time are usually more safe. |
| TRAVTIME | Distance to Work | Long drives to work usually suggest greater risk |
| URBANCITY | Home/Work Area | Unknown |
| YOJ | Years on Job | People who stay at a job for a long time are usually more safe |

# 1.2 Summary Statistics

Summary Statistics

| TARGET_FLAG | TARGET_AMT | KIDSDRIV | AGE | HOMEKIDS | YOJ |
|---|---|---|---|---|---|
| Min. :0.0000 | Min. : 0 | Min. :0.0000 | Min. :16.00 | Min. :0.0000 | Min. : 0.00 |
| 1st Qu.:0.0000 | 1st Qu.: 0 | 1st Qu.:0.0000 | 1st Qu.:39.00 | 1st Qu.:0.0000 | 1st Qu.: 9.00 |
| Median :0.0000 | Median : 0 | Median :0.0000 | Median :45.00 | Median :0.0000 | Median :11.00 |
| Mean :0.2638 | Mean : 1514 | Mean :0.1704 | Mean :44.68 | Mean :0.7298 | Mean :10.53 |
| 3rd Qu.:1.0000 | 3rd Qu.: 1029 | 3rd Qu.:0.0000 | 3rd Qu.:51.00 | 3rd Qu.:1.0000 | 3rd Qu.:13.00 |
| Max. :1.0000 | Max. :107586 | Max. :4.0000 | Max. :73.00 | Max. :5.0000 | Max. :19.00 |
| NA | NA | NA | NA's :3 | NA | NA's :347 |

Summary Statistics

| INCOME | PARENT1 | HOME_VAL | MSTATUS | SEX | EDUCATION | JOB |
|---|---|---|---|---|---|---|
| Min. : 0 | No :5289 | Min. : 0 | Yes :3659 | M :2827 | <High School : 933 | z_Blue Collar:1375 |
| 1st Qu.: 27658 | Yes: 832 | 1st Qu.: 0 | z_No:2462 | z_F:3294 | Bachelors :1693 | Clerical : 961 |
| Median : 53904 | NA | Median :159417 | NA | NA | Masters :1240 | Professional : 834 |
| Mean : 61890 | NA | Mean :154164 | NA | NA | PhD : 541 | Manager : 738 |
| 3rd Qu.: 86464 | NA | 3rd Qu.:238931 | NA | NA | z_High School:1714 | Lawyer : 615 |
| Max. :367030 | NA | Max. :885282 | NA | NA | NA | (Other) :1191 |
| NA's :343 | NA | NA's :344 | NA | NA | NA | NA's : 407 |

These tables give an overview of the variables, suggesting there may be some issues with distributions but we will need to look further before making any decisions on transforming the variables.

# 1.3 Descriptive Statistics

Descriptive Statistics

| | vars | n | mean | sd | median |
|---|---|---|---|---|---|
| TARGET_FLAG | 1 | 6121 | 2.638458e-01 | 4.407527e-01 | 0.0 |
| TARGET_AMT | 2 | 6121 | 1.514009e+03 | 4.776983e+03 | 0.0 |
| KIDSDRIV | 3 | 6121 | 1.703970e-01 | 5.102345e-01 | 0.0 |
| AGE | 4 | 6118 | 4.467980e+01 | 8.647718e+00 | 45.0 |
| HOMEKIDS | 5 | 6121 | 7.297827e-01 | 1.122196e+00 | 0.0 |
| YOJ | 6 | 5774 | 1.053152e+01 | 4.084161e+00 | 11.0 |
| INCOME | 7 | 5778 | 6.188946e+04 | 4.785552e+04 | 53903.5 |
| PARENT1* | 8 | 6121 | 1.135925e+00 | 3.427374e-01 | 1.0 |
| HOME_VAL | 9 | 5777 | 1.541642e+05 | 1.303360e+05 | 159417.0 |
| MSTATUS* | 10 | 6121 | 1.402222e+00 | 4.903863e-01 | 1.0 |
| SEX* | 11 | 6121 | 1.538147e+00 | 4.985834e-01 | 2.0 |
| EDUCATION* | 12 | 6121 | 3.066982e+00 | 1.445894e+00 | 3.0 |
| JOB* | 13 | 5714 | 5.008050e+00 | 2.468908e+00 | 5.0 |
| TRAVTIME | 14 | 6121 | 3.347100e+01 | 1.593407e+01 | 33.0 |
| CAR_USE* | 15 | 6121 | 1.627022e+00 | 4.836359e-01 | 2.0 |
| BLUEBOOK | 16 | 6121 | 1.567134e+04 | 8.455444e+03 | 14380.0 |
| TIF | 17 | 6121 | 5.349616e+00 | 4.106822e+00 | 4.0 |
| CAR_TYPE* | 18 | 6121 | 3.521157e+00 | 1.960455e+00 | 3.0 |
| RED_CAR* | 19 | 6121 | 1.290802e+00 | 4.541695e-01 | 1.0 |
| OLDCLAIM | 20 | 6121 | 4.070660e+03 | 8.867756e+03 | 0.0 |
| CLM_FREQ | 21 | 6121 | 7.952949e-01 | 1.152753e+00 | 0.0 |
| REVOKED* | 22 | 6121 | 1.121549e+00 | 3.267906e-01 | 1.0 |
| MVR_PTS | 23 | 6121 | 1.719164e+00 | 2.150401e+00 | 1.0 |
| CAR_AGE | 24 | 5735 | 8.334612e+00 | 5.747126e+00 | 8.0 |
| URBANICITY* | 25 | 6121 | 1.207156e+00 | 4.053012e-01 | 1.0 |

Descriptive Statistics

| | trimmed | mad | min | max |
|---|---|---|---|---|
| TARGET_FLAG | 2.048193e-01 | 0.0000 | 0 | 1.0 |
| TARGET_AMT | 5.889851e+02 | 0.0000 | 0 | 107586.1 |
| KIDSDRIV | 2.470900e-02 | 0.0000 | 0 | 4.0 |
| AGE | 4.473284e+01 | 8.8956 | 16 | 73.0 |
| HOMEKIDS | 5.064325e-01 | 0.0000 | 0 | 5.0 |
| YOJ | 1.110736e+01 | 2.9652 | 0 | 19.0 |
| INCOME | 5.676326e+04 | 42610.6653 | 0 | 367030.0 |
| PARENT1* | 1.044925e+00 | 0.0000 | 1 | 2.0 |
| HOME_VAL | 1.428207e+05 | 152887.1946 | 0 | 885282.0 |
| MSTATUS* | 1.377782e+00 | 0.0000 | 1 | 2.0 |
| SEX* | 1.547682e+00 | 0.0000 | 1 | 2.0 |
| EDUCATION* | 3.083725e+00 | 1.4826 | 1 | 5.0 |
| JOB* | 5.134952e+00 | 2.9652 | 1 | 8.0 |
| TRAVTIME | 3.301062e+01 | 16.3086 | 5 | 142.0 |
| CAR_USE* | 1.658771e+00 | 0.0000 | 1 | 2.0 |
| BLUEBOOK | 1.498274e+04 | 8510.1240 | 1500 | 69740.0 |
| TIF | 4.854605e+00 | 4.4478 | 1 | 25.0 |
| CAR_TYPE* | 3.526445e+00 | 2.9652 | 1 | 6.0 |
| RED_CAR* | 1.238513e+00 | 0.0000 | 1 | 2.0 |
| OLDCLAIM | 1.727992e+03 | 0.0000 | 0 | 57037.0 |
| CLM_FREQ | 5.860731e-01 | 0.0000 | 0 | 5.0 |
| REVOKED* | 1.026955e+00 | 0.0000 | 1 | 2.0 |
| MVR_PTS | 1.341842e+00 | 1.4826 | 0 | 13.0 |
| CAR_AGE | 7.954892e+00 | 7.4130 | 0 | 28.0 |
| URBANICITY* | 1.133960e+00 | 0.0000 | 1 | 2.0 |

Descriptive Statistics

| | range | skew | kurtosis | se |
|---|---|---|---|---|
| TARGET_FLAG | 1.0 | 1.0714201 | -0.8521981 | 0.0056336 |
| TARGET_AMT | 107586.1 | 8.7901683 | 115.2132560 | 61.0579864 |
| KIDSDRIV | 4.0 | 3.3505408 | 11.7603038 | 0.0065217 |
| AGE | 57.0 | -0.0564481 | -0.0854251 | 0.1105597 |
| HOMEKIDS | 5.0 | 1.3343569 | 0.6505442 | 0.0143436 |
| YOJ | 19.0 | -1.2180676 | 1.1808636 | 0.0537483 |
| INCOME | 367030.0 | 1.2145580 | 2.2994723 | 629.5688594 |
| PARENT1* | 1.0 | 2.1241626 | 2.5124772 | 0.0043808 |
| HOME_VAL | 885282.0 | 0.5294461 | 0.0815993 | 1714.7992717 |
| MSTATUS* | 1.0 | 0.3987149 | -1.8413272 | 0.0062680 |
| SEX* | 1.0 | -0.1529980 | -1.9769145 | 0.0063727 |
| EDUCATION* | 4.0 | 0.1363664 | -1.3715543 | 0.0184810 |
| JOB* | 7.0 | -0.3418602 | -1.1629758 | 0.0326614 |
| TRAVTIME | 137.0 | 0.4428075 | 0.7156313 | 0.2036646 |
| CAR_USE* | 1.0 | -0.5251925 | -1.7244545 | 0.0061817 |
| BLUEBOOK | 68240.0 | 0.8097543 | 0.8315495 | 108.0749994 |
| TIF | 24.0 | 0.8608451 | 0.3166601 | 0.0524922 |
| CAR_TYPE* | 5.0 | 0.0052126 | -1.5099694 | 0.0250580 |
| RED_CAR* | 1.0 | 0.9210819 | -1.1517962 | 0.0058051 |
| OLDCLAIM | 57037.0 | 3.1274684 | 9.9061482 | 113.3450445 |

| | | | | |
|---|---|---|---|---|
| CLM_FREQ | 5.0 | 1.2115436 | 0.3061895 | 0.0147342 |
| REVOKED* | 1.0 | 2.3157911 | 3.3634379 | 0.0041769 |
| MVR_PTS | 13.0 | 1.3131263 | 1.2250866 | 0.0274858 |
| CAR_AGE | 28.0 | 0.2984052 | -0.7372102 | 0.0758899 |
| URBANICITY* | 1.0 | 1.4448340 | 0.0875597 | 0.0051804 |

The count of NA values for each variable is given below.

Count of NA Values

| x | x | x | x | x | | x | | x | x |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 3 | 0 | | 347 | | 343 | 0 |

Count of NA Values

| | x | x | x | x | | x | x | x | x | x |
|---|---|---|---|---|---|---|---|---|---|---|
| 344 | 0 | 0 | 0 | | 407 | 0 | 0 | 0 | 0 | |

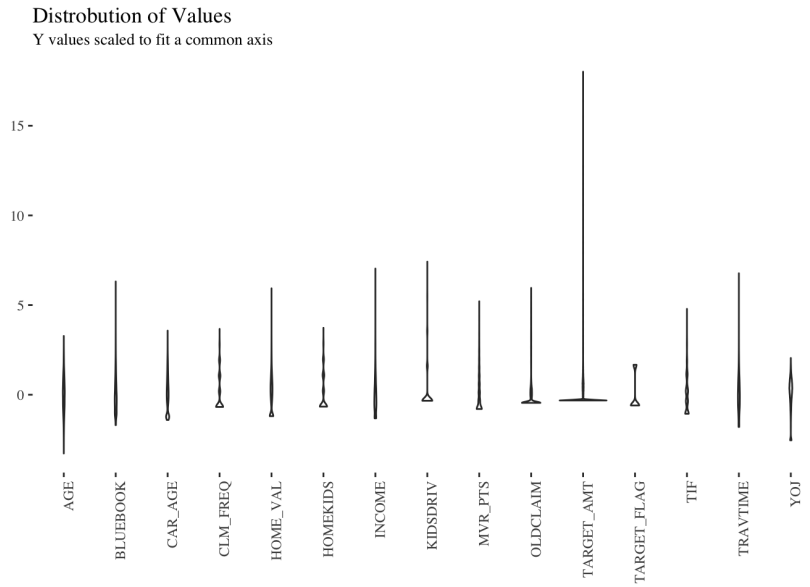
Count of NA Values

| x | x | x | x | x | x | | x | x |
|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 0 | 0 | 0 | 386 | 0 | |

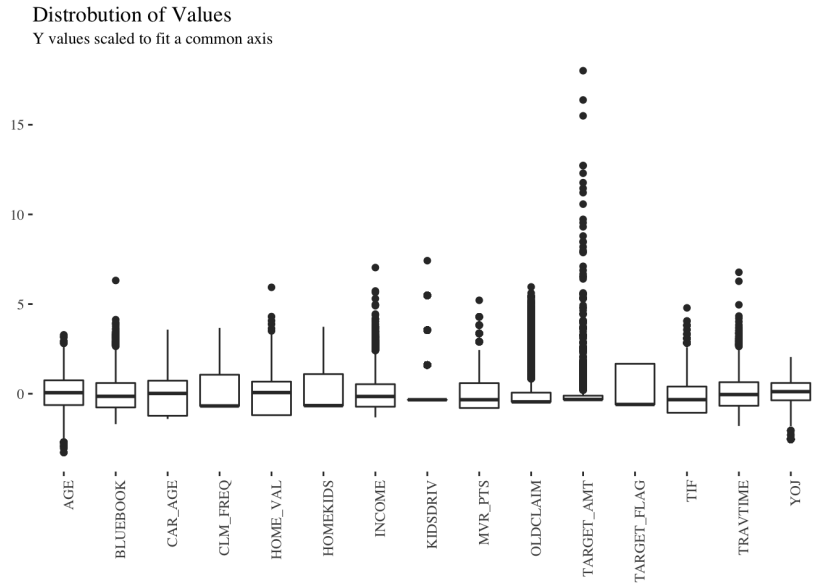
There are quite a few missing values accross several variables. However, compared to the size of the training set, around 6000, these numbers could be dropped if there is no correlation between the missing values and the response variables.

The correlation between missing values and the ‘Claim Filed’ response is -3.151689910^{-4} and 0.0086636 for the claim amount. Since these are very close to zero we are not worried about them effecting the regressions. As such, we will drop them.

# 1.4 Graphical EDA

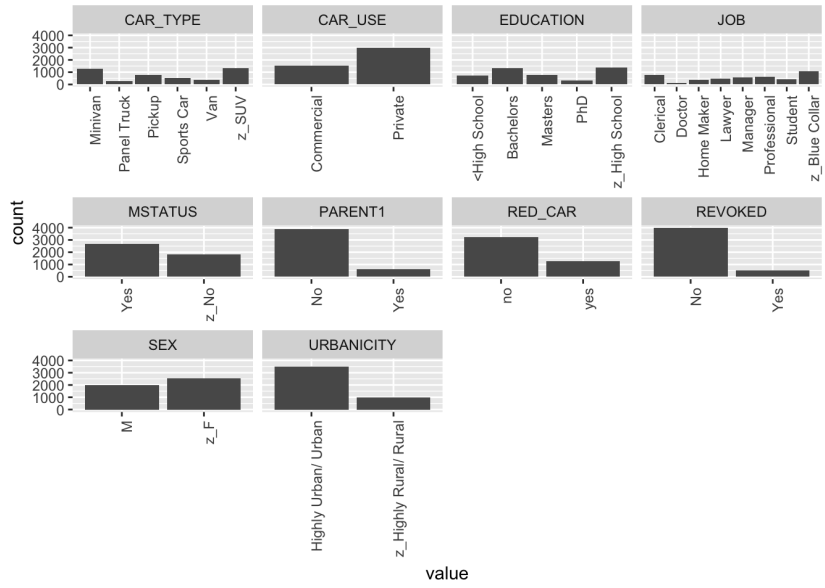



The distrobutions are generally skeded upwards, nothing suggests problems with the dataset. The only variable that is very skewed is `TARGET_AMT` and this makes sense because most are zero or low and some are very high.




From these two graphs we can see that many of distributions are skewed in one direction or another. It is also interesting to see that the target variable is below zero. This means that the median and mean values are different.

```
## Warning: attributes are not identical across measure variables;
## they will be dropped
```




The factor variables some uneven counts as well but nothing that is highly out of the ordinary.

# 2 Data Preperation

## 2.1 Transformed Skewed Variables

I will log transform `TARGET_AMT` in one of the models that build to account for the wide range. During this transformation it is important to add 1 to each variable because there are many zero values that would throw and error.

I also transformed `TARGET_AMT` with the power of -0.4 based off of a Box-Cox analysis. Given the nature of other variables it does not seem necessary to transform others.

# 3 Build Models

Model selection will be based off of automated selection techniques as well as specific transformations.

## 3.1 Ordinary Least Squares

### 3.1.1 Baisic Regression

Here I regressed all variables without transformation.

Call: lm(formula = TARGET_AMT ~ ., data = dfCont)

Residuals: Min 1Q Median 3Q Max -5602 -1702 -733 419 82515

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.206e+03 6.267e+02 3.520 0.000436 ***KIDSDRIV 2.624e+02 1.487e+02 1.765 0.077705 .***
***AGE -5.843e+00 9.387e+00 -0.622 0.533648***
***HOMEKIDS -3.457e+01 8.506e+01 -0.406 0.684471***
***YOJ -8.626e+00 1.944e+01 -0.444 0.657279***
***INCOME -2.588e-03 2.719e-03 -0.952 0.341152***
***PARENT1Yes 6.711e+02 2.643e+02 2.539 0.011153***
**HOME_VAL -1.373e-03 8.504e-04 -1.615 0.106471**
**MSTATUSz_No 4.870e+02 2.001e+02 2.434 0.014982 ***
**SEXz_F -3.404e+02 2.408e+02 -1.414 0.157515**
**EDUCATIONBachelors -3.255e+02 2.662e+02 -1.222 0.221603**
**EDUCATIONMasters -3.557e+02 4.018e+02 -0.885 0.376129**
**EDUCATIONPhD 5.905e+02 4.993e+02 1.183 0.236960**
**EDUCATIONz_High School -1.534e+02 2.200e+02 -0.697 0.485834**
**JOBDoctor -1.365e+03 5.902e+02 -2.313 0.020762 ***
**JOBHome Maker 1.498e+01 3.263e+02 0.046 0.963375**
**JOBLawyer 1.771e+02 4.001e+02 0.443 0.658116**
**JOBManager -9.372e+02 3.075e+02 -3.048 0.002320** JOBProfessional 2.831e+02 2.825e+02 1.002 0.316268
JOBStudent -2.023e+02 3.145e+02 -0.643 0.520122
JOBz_Blue Collar 8.733e+01 2.464e+02 0.354 0.723062
TRAVTIME 1.666e+01 4.277e+00 3.895 9.98e-05 ***CAR_USEPrivate -8.516e+02 2.169e+02 -3.926 8.76e-05*** BLUEBOOK 1.124e-02 1.139e-02 0.987 0.323848
TIF -4.546e+01 1.639e+01 -2.773 0.005570 ** CAR_TYPEPanel Truck 3.648e+02 3.928e+02 0.929 0.353128
CAR_TYPEPickup 4.390e+02 2.222e+02 1.975 0.048285 *
CAR_TYPESports Car 1.304e+03 2.773e+02 4.704 2.63e-06 ***CAR_TYPEVan 5.659e+02 2.912e+02 1.943 0.052050 .***
***CAR_TYPEz_SUV 7.651e+02 2.299e+02 3.328 0.000881*** RED_CARyes -1.374e+02 2.034e+02 -0.676 0.499234
OLDCLAIM -7.776e-03 9.851e-03 -0.789 0.429963
CLM_FREQ 5.019e+01 7.386e+01 0.680 0.496839
REVOKEDYes 3.721e+02 2.356e+02 1.580 0.114251
MVR_PTS 1.654e+02 3.434e+01 4.817 1.51e-06 ***CAR_AGE -2.414e+01 1.720e+01 -1.404 0.160504***
***URBANICITYz_Highly Rural/ Rural -1.918e+03 1.814e+02 -10.573 < 2e-16*** — Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4494 on 4468 degrees of freedom Multiple R-squared: 0.07951, Adjusted R-squared: 0.07209 F-statistic: 10.72 on 36 and 4468 DF, p-value: < 2.2e-16

This is a pretty poor $R^2$. While there are some significant variables, the overall performance is poor.

### 3.1.2 Log Transformed

Next we look at the regression with a log transformed `TARGET_AMT` variable.

Call: lm(formula = log(TARGET_AMT + 1) ~ ., data = dfCont)

Residuals: Min 1Q Median 3Q Max -8.173 -2.340 -0.850 2.146 11.054

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 3.640e+00 4.466e-01 8.150 4.66e-16 ***KIDSDRIV 3.949e-01 1.060e-01 3.727 0.000197*** AGE -6.131e-03 6.689e-03 -0.916 0.359457
HOMEKIDS -2.705e-02 6.061e-02 -0.446 0.655468
YOJ -1.264e-02 1.385e-02 -0.912 0.361683
INCOME -3.189e-06 1.937e-06 -1.646 0.099790 .
PARENT1Yes 8.902e-01 1.884e-01 4.726 2.36e-06 ***HOME_VAL -1.442e-06 6.060e-07 -2.379 0.017379***
**MSTATUSz_No 4.756e-01 1.426e-01 3.335 0.000860** *SEXz_F -1.547e-01 1.716e-01 -0.902 0.367280*
*EDUCATIONBachelors -5.226e-01 1.897e-01 -2.755 0.005901* ***EDUCATIONMasters -4.859e-01 2.864e-01 -1.697 0.089781 .***
***EDUCATIONPhD 1.498e-01 3.558e-01 0.421 0.673785***
***EDUCATIONz_High School 4.030e-02 1.568e-01 0.257 0.797156***
***JOBDoctor -1.388e+00 4.206e-01 -3.299 0.000978*** JOBHome Maker -3.566e-01 2.325e-01 -1.534 0.125162
JOBLawyer -2.596e-01 2.851e-01 -0.910 0.362667
JOBManager -1.395e+00 2.191e-01 -6.368 2.11e-10 ***JOBProfessional -4.193e-01 2.013e-01 -2.083 0.037331***
**JOBStudent -4.121e-01 2.241e-01 -1.839 0.065963 .**
**JOBz_Blue Collar -3.573e-01 1.756e-01 -2.035 0.041932 ***
**TRAVTIME 2.163e-02 3.048e-03 7.095 1.50e-12** *CAR_USEPrivate -1.199e+00 1.546e-01 -7.754 1.10e-14* **BLUEBOOK -2.055e-05 8.115e-06 -2.532 0.011368 ***
**TIF -6.686e-02 1.168e-02 -5.724 1.11e-08** *CAR_TYPEPanel Truck 5.446e-01 2.800e-01 1.945 0.051794 .*
*CAR_TYPEPickup 6.586e-01 1.584e-01 4.159 3.26e-05* **CAR_TYPESports Car 1.332e+00 1.976e-01 6.740 1.78e-11** *CAR_TYPEVan 5.699e-01 2.075e-01 2.746 0.006057* ***CAR_TYPEz_SUV 8.588e-01 1.638e-01 5.242 1.66e-07*** RED_CARyes -2.352e-01 1.449e-01 -1.623 0.104640
OLDCLAIM -1.883e-05 7.020e-06 -2.682 0.007347 ** CLM_FREQ 2.477e-01 5.263e-02 4.706 2.60e-06 ***REVOKEDYes 1.058e+00 1.679e-01 6.300 3.27e-10*** MVR_PTS 1.949e-01 2.447e-02 7.965 2.07e-15 ***CAR_AGE -8.615e-03 1.226e-02 -0.703 0.482150***
***URBANICITYz_Highly Rural/ Rural -2.571e+00 1.293e-01 -19.894 < 2e-16*** — Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.203 on 4468 degrees of freedom Multiple R-squared: 0.2429, Adjusted R-squared: 0.2368 F-statistic: 39.81 on 36 and 4468 DF, p-value: < 2.2e-16

This is an improvement. It makes sense that we would need to transform the response variable given it’s skewed nature.

## 3.2 BoxCox

bcPower Transformation to Normality Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd Y1 -0.3997 -0.4 -0.4151 -0.3842

Likelihood ratio tests about transformation parameters LRT df pval LR test, lambda = (0) 3334.454 1 0 LR test, lambda = (1) 48980.134 1 0

Call: lm(formula = I((TARGET_AMT + 1)^(-0.4)) ~ ., data = dfCont)

Residuals: Min 1Q Median 3Q Max -1.22372 -0.26618 0.09885 0.27222 0.94841

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 5.718e-01 5.158e-02 11.087 < 2e-16 ***KIDSDRIV -4.597e-02 1.224e-02 -3.756 0.000175*** AGE 7.686e-04 7.725e-04 0.995 0.319821
HOMEKIDS 3.092e-03 7.000e-03 0.442 0.658701
YOJ 1.362e-03 1.600e-03 0.851 0.394739
INCOME 3.641e-07 2.238e-07 1.627 0.103780
PARENT1Yes -1.008e-01 2.175e-02 -4.633 3.71e-06 ***HOME_VAL 1.613e-07 6.999e-08 2.304 0.021244***
**MSTATUSz_No -5.537e-02 1.647e-02 -3.362 0.000781** *SEXz_F 1.657e-02 1.982e-02 0.836 0.403263*
*EDUCATIONBachelors 5.935e-02 2.191e-02 2.709 0.006782* ***EDUCATIONMasters 5.585e-02 3.307e-02 1.689 0.091322 .***
***EDUCATIONPhD -1.270e-02 4.109e-02 -0.309 0.757357***
***EDUCATIONz_High School -4.987e-03 1.811e-02 -0.275 0.783029***
***JOBDoctor 1.581e-01 4.858e-02 3.255 0.001140*** *JOBHome Maker 4.226e-02 2.686e-02 1.573 0.115690*
*JOBLawyer 3.300e-02 3.293e-02 1.002 0.316351*
*JOBManager 1.629e-01 2.531e-02 6.438 1.34e-10* **JOBProfessional 5.427e-02 2.325e-02 2.334 0.019625 ***
**JOBStudent 4.755e-02 2.588e-02 1.837 0.066220 .**
**JOBz_Blue Collar 4.227e-02 2.028e-02 2.084 0.037178 ***
**TRAVTIME -2.493e-03 3.520e-04 -7.080 1.66e-12** *CAR_USEPrivate 1.393e-01 1.785e-02 7.801 7.62e-15* **BLUEBOOK 2.603e-06 9.372e-07 2.778 0.005501** TIF 7.735e-03 1.349e-03 5.734 1.05e-08 ***CAR_TYPEPanel Truck -6.329e-02 3.233e-02 -1.957 0.050359 .***
***CAR_TYPEPickup -7.585e-02 1.829e-02 -4.147 3.43e-05*** CAR_TYPESports Car -1.518e-01 2.282e-02 -6.650 3.29e-11 ***CAR_TYPEVan -6.372e-02 2.397e-02 -2.659 0.007873*** *CAR_TYPEz_SUV -9.798e-02 1.892e-02 -5.179 2.33e-07* **RED_CARyes 2.835e-02 1.674e-02 1.694 0.090340 .**
**OLDCLAIM 2.323e-06 8.108e-07 2.865 0.004186** CLM_FREQ -3.064e-02 6.079e-03 -5.041 4.82e-07 ***REVOKEDYes -1.255e-01 1.939e-02 -6.471 1.08e-10*** MVR_PTS -2.202e-02 2.826e-03 -7.791 8.19e-15 ***CAR_AGE 9.464e-04 1.415e-03 0.669 0.503748***
***URBANICITYz_Highly Rural/ Rural 2.970e-01 1.493e-02 19.898 < 2e-16*** — Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3699 on 4468 degrees of freedom Multiple R-squared: 0.2444, Adjusted R-squared: 0.2383 F-statistic: 40.13 on 36 and 4468 DF, p-value: < 2.2e-16

# 3.3 Regular logostic

Call: glm(formula = TARGET_FLAG ~ ., data = dfLog)

Deviance Residuals: Min 1Q Median 3Q Max
-0.9859 -0.2831 -0.1032 0.2783 1.2625

Coefficients: Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.462e-01 5.365e-02 8.316 < 2e-16 ***KIDSDRIV 4.778e-02 1.273e-02 3.753 0.000177*** AGE -8.077e-04 8.036e-04 -1.005 0.314936
HOMEKIDS -3.191e-03 7.282e-03 -0.438 0.661319
YOJ -1.397e-03 1.665e-03 -0.839 0.401432
INCOME -3.775e-07 2.328e-07 -1.622 0.104895
PARENT1Yes 1.043e-01 2.263e-02 4.607 4.19e-06 ***HOME_VAL -1.665e-07 7.281e-08 -2.286 0.022276***
**MSTATUSz_No 5.775e-02 1.713e-02 3.370 0.000757** *SEXz_F -1.718e-02 2.062e-02 -0.833 0.404608*
*EDUCATIONBachelors -6.133e-02 2.279e-02 -2.690 0.007162* ***EDUCATIONMasters -5.796e-02 3.440e-02 -1.685 0.092145 .***
***EDUCATIONPhD 1.286e-02 4.274e-02 0.301 0.763509***
***EDUCATIONz_High School 5.205e-03 1.884e-02 0.276 0.782319***
***JOBDoctor -1.644e-01 5.053e-02 -3.253 0.001149*** *JOBHome Maker -4.396e-02 2.794e-02 -1.574 0.115658*
*JOBLawyer -3.464e-02 3.426e-02 -1.011 0.311951*
*JOBManager -1.696e-01 2.633e-02 -6.444 1.29e-10* **JOBProfessional -5.725e-02 2.419e-02 -2.367 0.017962 ***
**JOBStudent -4.933e-02 2.692e-02 -1.832 0.066990 .**
**JOBz_Blue Collar -4.370e-02 2.110e-02 -2.071 0.038385 ***
**TRAVTIME 2.589e-03 3.662e-04 7.071 1.78e-12** *CAR_USEPrivate -1.449e-01 1.857e-02 -7.802 7.52e-15* **BLUEBOOK -2.743e-06 9.749e-07 -2.814 0.004916** TIF -8.044e-03 1.403e-03 -5.732 1.06e-08 ***CAR_TYPEPanel Truck 6.574e-02 3.363e-02 1.955 0.050694 .***
***CAR_TYPEPickup 7.887e-02 1.903e-02 4.146 3.45e-05*** CAR_TYPESports Car 1.575e-01 2.374e-02 6.632 3.71e-11 ***CAR_TYPEVan 6.581e-02 2.493e-02 2.640 0.008331*** *CAR_TYPEz_SUV 1.015e-01 1.968e-02 5.159 2.59e-07* **RED_CARyes -2.978e-02 1.741e-02 -1.710 0.087280 .**
**OLDCLAIM -2.447e-06 8.434e-07 -2.901 0.003740** CLM_FREQ 3.222e-02 6.324e-03 5.096 3.62e-07 ***REVOKEDYes 1.308e-01 2.017e-02 6.484 9.92e-11*** MVR_PTS 2.279e-02 2.940e-03 7.752 1.11e-14 ***CAR_AGE -9.862e-04 1.472e-03 -0.670 0.503048***
***URBANICITYz_Highly Rural/ Rural -3.087e-01 1.553e-02 -19.879 < 2e-16*** — Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 0.1480396)

Null deviance: 875.19 on 4504 degrees of freedom Residual deviance: 661.44 on 4468 degrees of freedom AIC: 4217.7

Number of Fisher Scoring iterations: 2

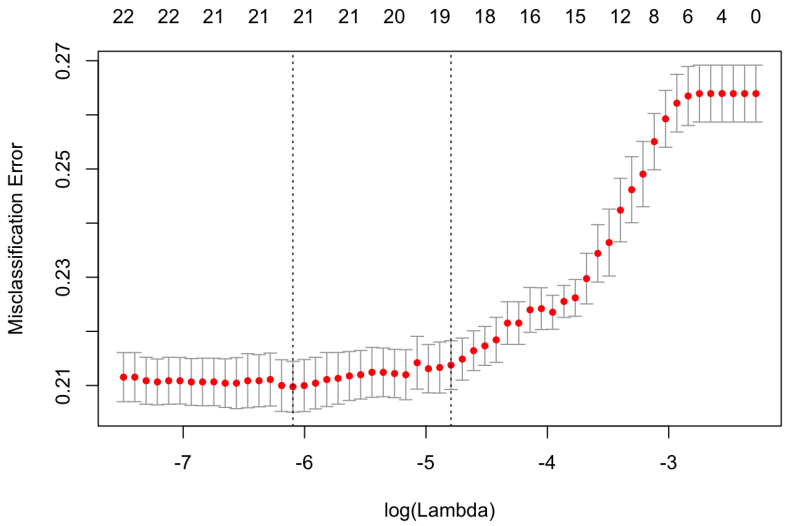
# 3.4 LASSO Logistic

```
## Loading required package: gplots
```

```
##
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
##
##     lowess
```

```
## Warning in aucDF$lasso.prob <- predict(lasso.model, type = "response", newx
## = X, : Coercing LHS to a list
```




```
## 24 x 1 sparse Matrix of class "dgCMatrix"
##                         1
## (Intercept)  9.311645e-01
## KIDSDRIV     2.104515e-01
## AGE         -3.066729e-03
## HOMEKIDS     .
## YOJ          .
## INCOME      -5.091546e-06
## PARENT1      5.400654e-01
## HOME_VAL    -1.319478e-06
## MSTATUS      2.495976e-01
## SEX          .
## EDUCATION    1.950305e-02
## JOB         -3.135790e-04
## TRAVTIME     1.298654e-02
## CAR_USE     -8.073200e-01
## BLUEBOOK    -1.739624e-05
## TIF         -3.631229e-02
## CAR_TYPE     9.955364e-02
## RED_CAR     -3.785589e-02
## OLDCLAIM     .
## CLM_FREQ     1.374393e-01
## REVOKED      4.762156e-01
## MVR_PTS      1.130536e-01
## CAR_AGE     -1.716225e-02
## URBANICITY  -1.888708e+00
```

### 3.4.1 Theoretical Model

This model is selected for variables that I think will play a larger role based on my prior belifes.

# 4 Chose a Model

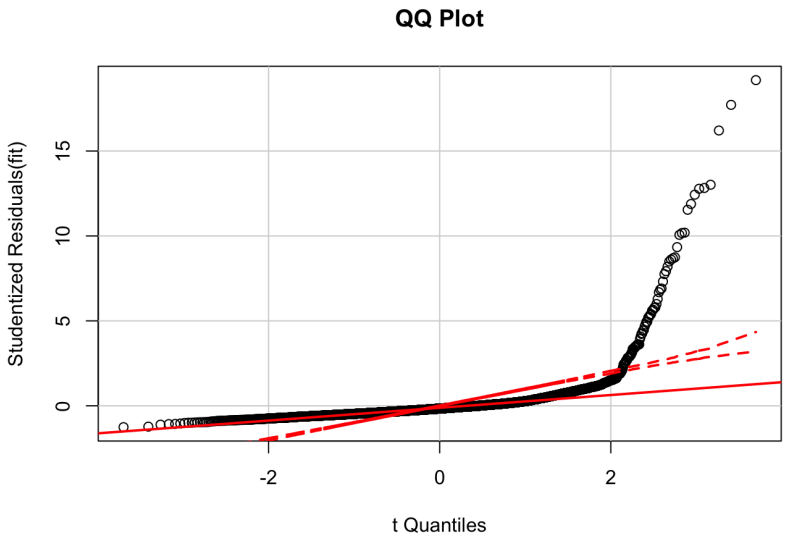
Here we chose the best of each class of model.

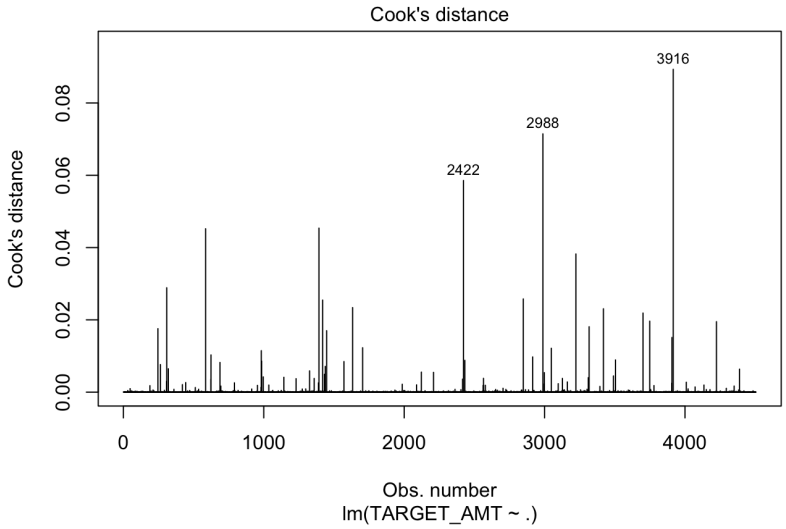
## 4.1 OLS Models

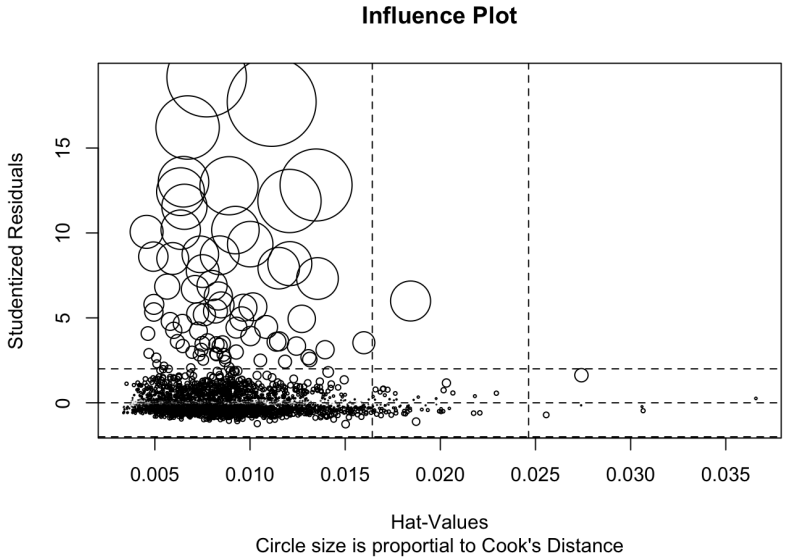
### 4.1.1 Regular OLS

```
##      rstudent unadjusted p-value Bonferonni p
## 2988 19.17386         8.3009e-79   3.7395e-75
## 3916 17.71827         6.0975e-68   2.7469e-64
## 1393 16.20840         1.8631e-57   8.3935e-54
## 309  13.01407         4.9631e-38   2.2359e-34
## 2422 12.82457         5.3304e-37   2.4013e-33
## 3223 12.77978         9.2987e-37   4.1891e-33
## 2848 12.43057         6.6975e-35   3.0172e-31
## 586  11.88084         4.5067e-32   2.0303e-28
## 3419 11.54189         2.1799e-30   9.8206e-27
## 246  10.19349         3.8883e-24   1.7517e-20
```

```
## Warning in rlm.default(x, y, weights, method = method, wt.method =
## wt.method, : 'rlm' failed to converge in 20 steps
```

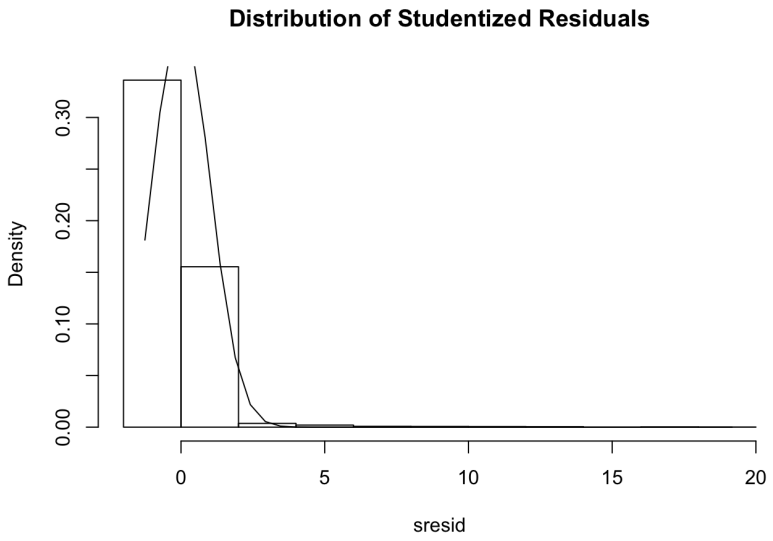


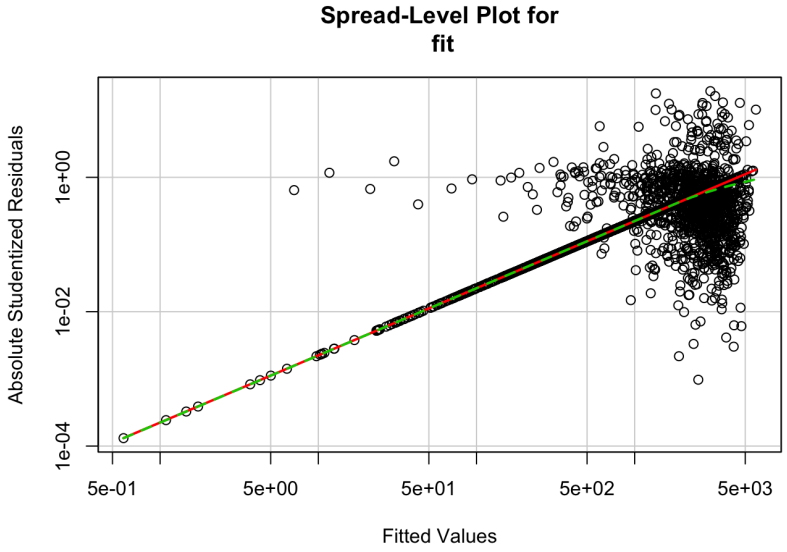




```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1937.732    Df = 1     p = 0
```

```
## Warning in spreadLevelPlot.lm(fit): 631 negative fitted values removed
```



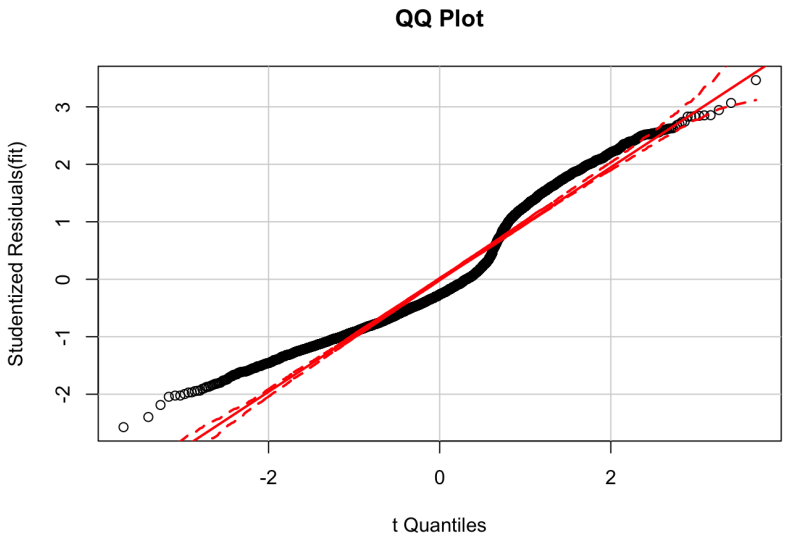



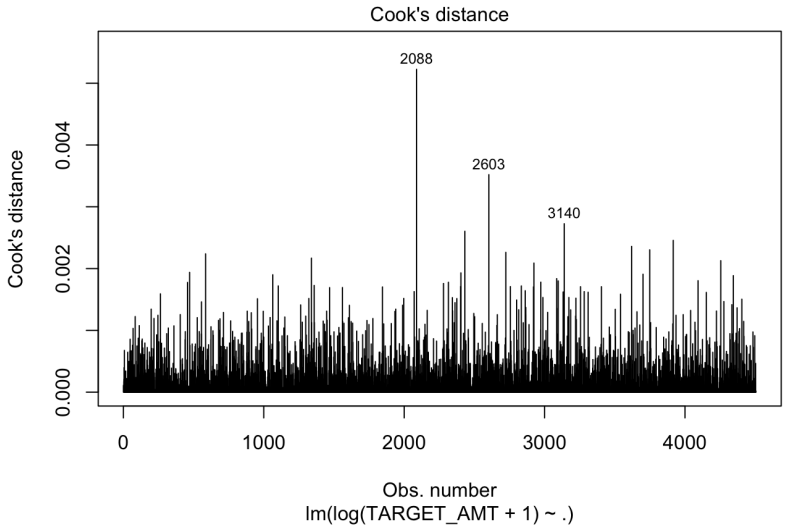

```
##  lag Autocorrelation D-W Statistic p-value
##    1    -0.003806213      2.007557   0.892
##  Alternative hypothesis: rho != 0
```

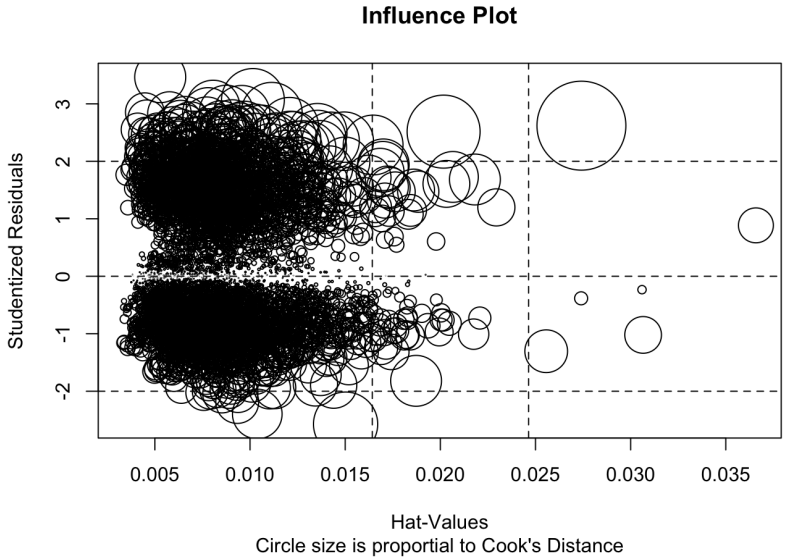
We can see from the QQ Plot that we have some serious troubles with this model.

Let's hope that we can find someting better. ### Log Scaled Model

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferonni p
## 2834 3.464935         0.00053535           NA
```

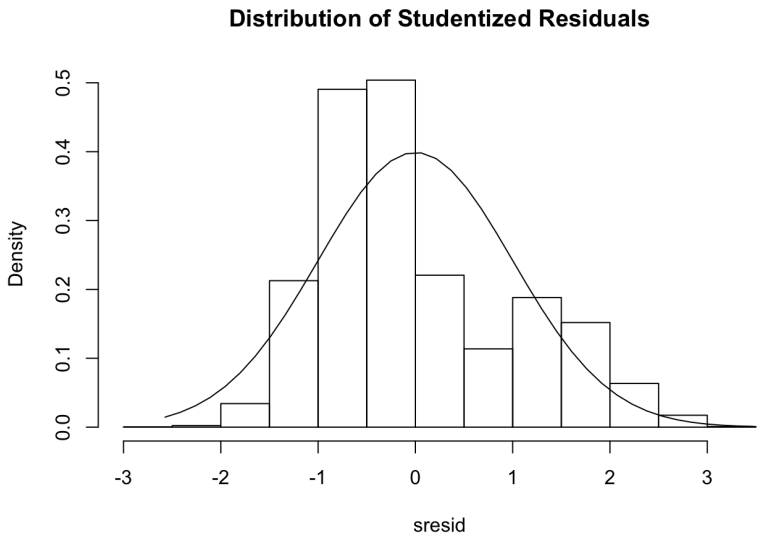


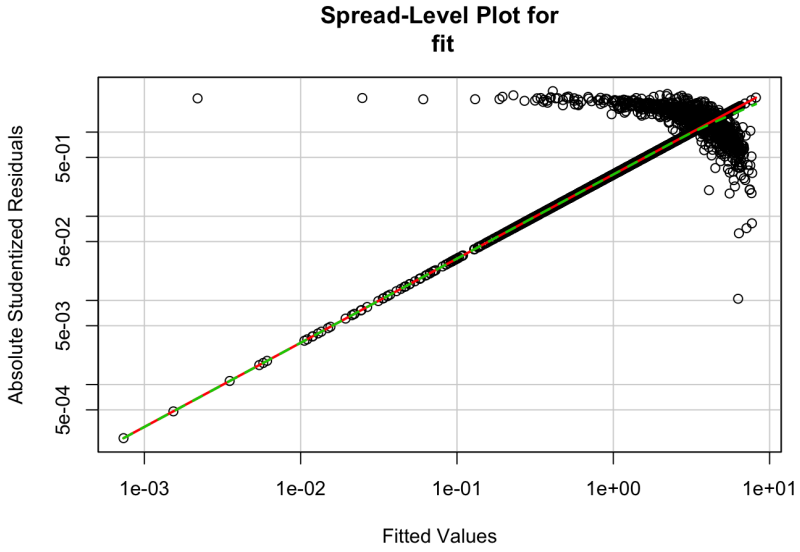




```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 419.1293    Df = 1     p = 3.77547e-93
```

```
## Warning in spreadLevelPlot.lm(fit): 534 negative fitted values removed
```

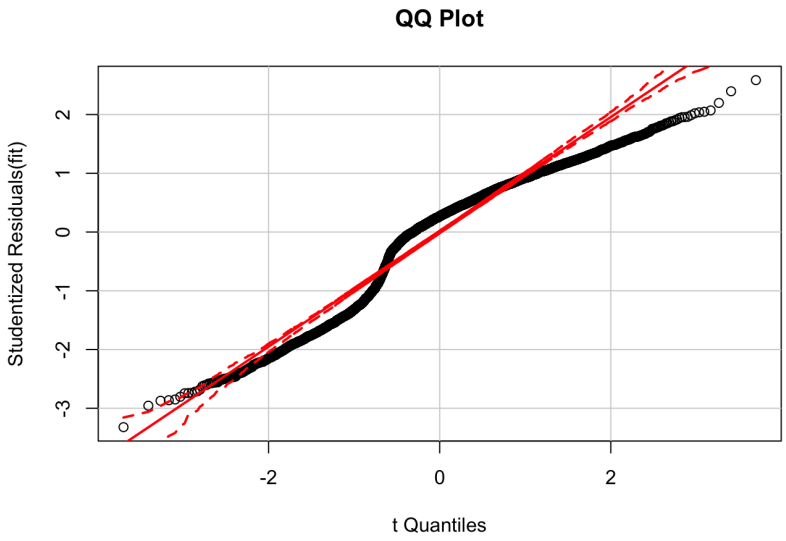


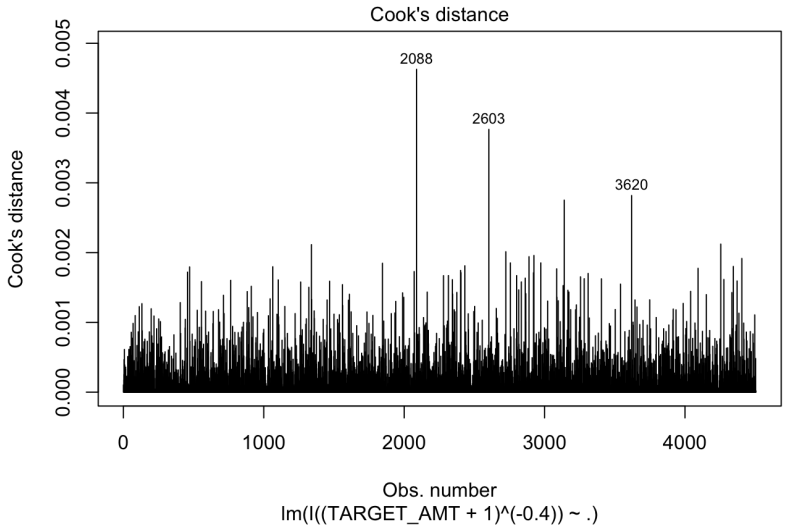



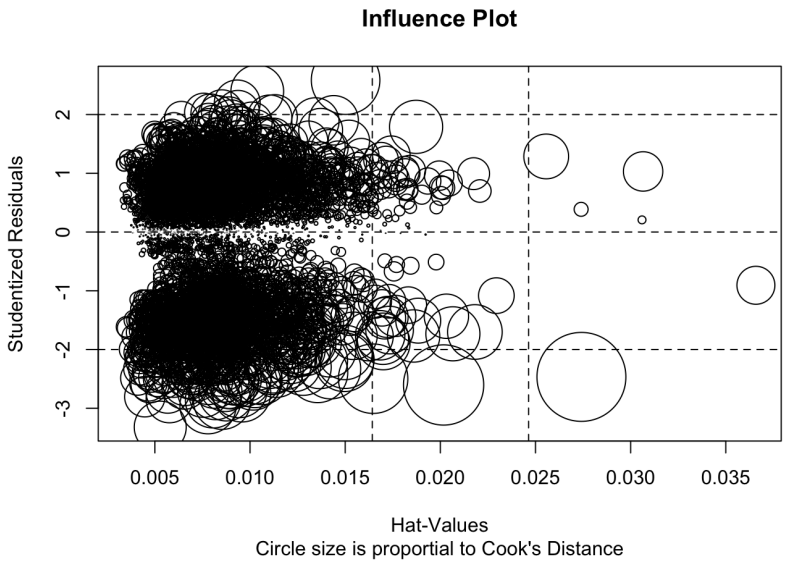

```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.02836693      2.056647   0.038
##  Alternative hypothesis: rho != 0
```

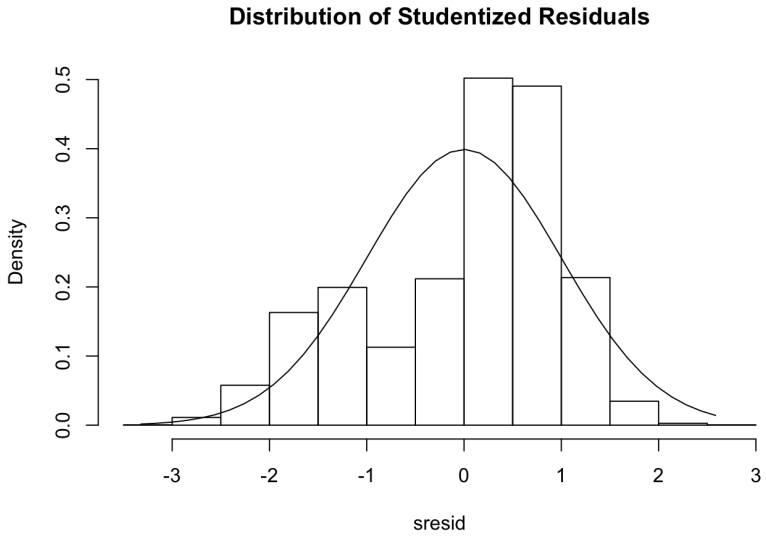
This model looks better but it is still far from perfect. The QQ Plot is greatly improved but there are still issues with cooks distance etc.

```
##
## No Studentized residuals with Bonferonni p < 0.05
## Largest |rstudent|:
##          rstudent unadjusted p-value Bonferonni p
## 2834 -3.32102....         0.00090407           NA
```

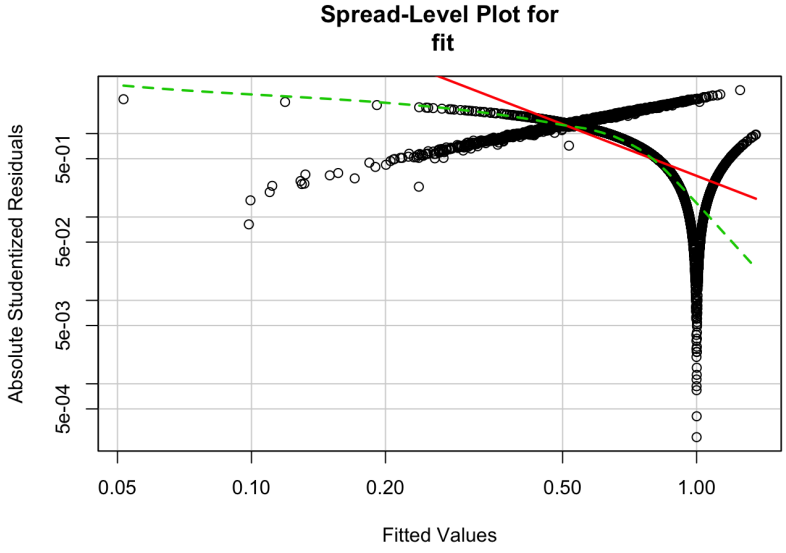

QQ Plot
Studentized Residuals(fit)
-3
-2
-1
0
1
2
-2
0
2
t Quantiles


Cook's distance
2088
2603
3620
Cook's distance
0.000
0.001
0.002
0.003
0.004
0.005
0
1000
2000
3000
4000
Obs. number
lm(I((TARGET_AMT + 1)^(-0.4)) ~ .)




```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 394.5613    Df = 1     p = 8.411692e-88
```




```
##  lag Autocorrelation D-W Statistic p-value
##    1     -0.03132353      2.062565   0.018
```

```
##  Alternative hypothesis: rho != 0
```

Even this model with a Box-Cox transformation has not great results. Given these results on our transformed models, I think that it might be worthwile examaning non-linear models such as tree based models. This, however, goes beyond the scope of the course.

# 4.2 Logisic Models

## 4.2.1 Basic Logistic regression

```
## Warning: package 'pROC' was built under R version 3.4.4
```

```
## Type 'citation("pROC")' for a citation.
```

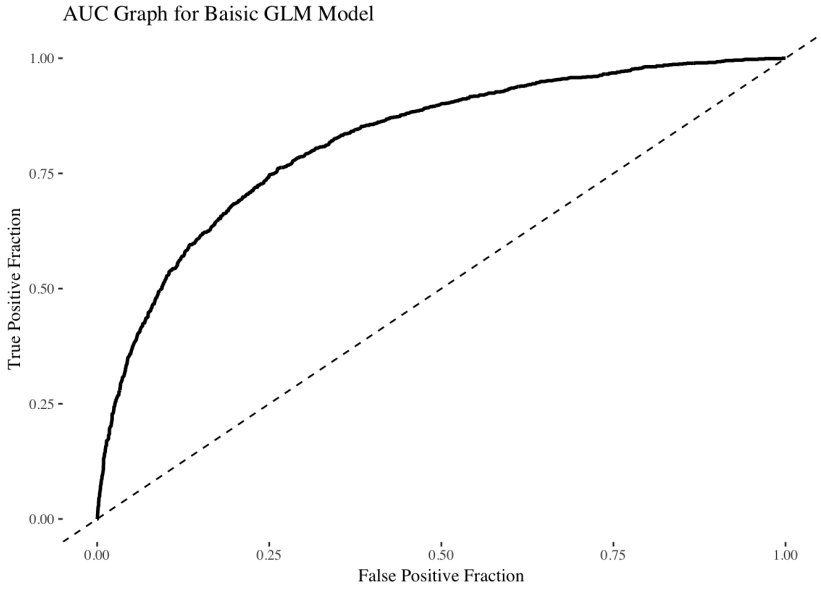
```
##
## Attaching package: 'pROC'
```

```
## The following object is masked from 'package:plotROC':
##
##     ggroc
```

```
## The following object is masked from 'package:glmnet':
##
##     auc
```

```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

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

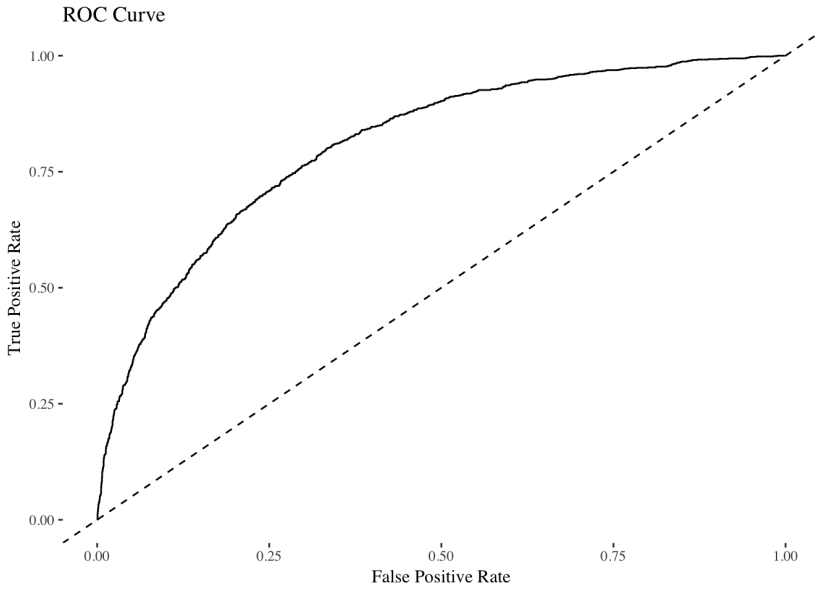



```
##
## [[2]]
## Area under the curve: 0.8207
```

## 4.2.2

The theoretical model performs much worse than the model with everyting in it.

## 4.2.3 LASSO Logsitic




```
## [[1]]
## [1] 0.8087627
```

# 5 Model Selection

## 5.1 OLS

I will base my selection off of the $R^2$ value. As such, we select the log transfomred logistic regression.

## 5.2 GLM Selection.

For the GLM Model we will use AUC as a selector. As such, I chose the baisc model.

# 6 Predictions

## 6.1 OLS

```
##          1          2          3          4          5          6
## 1.07923964 2.50448156 1.98003329 1.54566376 3.63463698 0.05468482
```

## 6.2 GLM

```
##          1          2          3          4          5          6
## 0.12471795 0.29952584 0.24145782 0.17842149 0.43940253 0.01132441
```

# 7 Apendix

```
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_knit$set(root.dir = "/Users/kailukowiak/DATA621/Assignments")
knitr::opts_chunk$set(echo=FALSE)
# Libraries
###################
library(MASS)
library(car)
library(leaps)
library(tidyverse)
library(knitr)
library(kableExtra)
library(psych)
library(ggthemes)
library(corrplot)
library(glmnet)
library(bestglm)
library(xtable)
library(caTools)
options(xtable.floating = FALSE)
options(xtable.timestamp = "")
####################
moneyCV <- function(df){
  for (colName in names(df)){
    if (grepl('\\$', df[, colName])){
```

```
      df[, colName] = gsub("\\$|,", "", df[[colName]]) %>% as.numeric()
    }
  }
  return(df)
}

factorCV <- function(df){
  for (colName in names(df)){
    if (is.character(df[[colName]])) {
      df[, colName] = df[[colName]] %>% as.factor()
    }
  }
  return(df)
}
# Loading the data

LabledDF <- read_csv('Assignment4/insurance_training_data.csv')

LabledDF <- moneyCV(LabledDF)
LabledDF <- factorCV(LabledDF)
LabledDF <- LabledDF %>% select(-INDEX)
set.seed(101)
sample = sample.split(LabledDF$TARGET_FLAG, SplitRatio = .75)
df <- subset(LabledDF, sample == TRUE)
testDF <- subset(LabledDF, sample == FALSE)
temp <- df %>% sample_n(6)
temp %>% kable()
# temp[,1:8] %>% kable(caption = 'Sample of Values for the Training Set')
# temp[9:17] %>% kable(caption = 'Sample of Values for the Training Set')
# temp[18:25] %>% kable(caption = 'Sample of Values for the Training Set')
glimpse(df)
evalDF <- read_csv('/Users/kailukowiak/DATA621/Assignments/Assignment4/insurance-evaluation-data.csv')
evalDF <- moneyCV(evalDF)
evalDF <- factorCV(evalDF)
evalDF <- evalDF %>% select(-INDEX, -TARGET_AMT, -TARGET_FLAG)
temp <- evalDF %>% sample_n(6)
temp[1:8] %>% kable(caption = 'Sample of Values for the Test Set')
################################
setwd("~/DATA621/Assignments")
lables <- read_csv('Assignment4/dataLegend.csv')
lables %>% kable()
# Summary Tables
SumTab <- summary(df)
SumTab1 <- SumTab[, 1:6]
SumTab2 <- SumTab[, 7:13]
kable(SumTab1, caption = 'Summary Statistics')
kable(SumTab2, caption = 'Summary Statistics')
######################
dis <- describe(df)
dis[, 1:5] %>% kable(caption = 'Descriptive Statistics')
dis[, 6:9] %>% kable(caption = 'Descriptive Statistics')
dis[, 10:13] %>% kable(caption = 'Descriptive Statistics')
temp <- map(df, ~sum(is.na(.)))
temp <- t(temp)
temp[1:8] %>% kable(caption = 'Count of NA Values')
temp[9:17] %>% kable(caption = 'Count of NA Values')
temp[18:25] %>% kable(caption = 'Count of NA Values')

df$CONTAINS_NA <- ifelse(complete.cases(df), FALSE, TRUE)

corFlag <- cor(df$TARGET_FLAG, df$CONTAINS_NA)
corAmt <- cor(df$TARGET_AMT, df$CONTAINS_NA)
df <- df %>% select(-CONTAINS_NA)
df <- df[complete.cases(df),]
evalDF <- evalDF[complete.cases(evalDF),]
testDF <- testDF[complete.cases(testDF),]
df %>%
  select_if(is.numeric) %>%
  scale() %>%
  as_tibble() %>%
  gather() %>%
  ggplot(aes(x = key, y = value)) +
  theme_tufte() +
  geom_violin()+
  #geom_tufteboxplot(outlier.colour="black")+
  theme(axis.title=element_blank()) +
  ylab('Scaled Values')+
  xlab('Variable')+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggtitle('Distrobution of Values', subtitle = 'Y values scaled to fit a common axis')
df %>%
  select_if(is.numeric) %>%
  scale() %>%
  as_tibble() %>%
  gather() %>%
  ggplot(aes(x = key, y = value)) +
  # geom_violin()+
  # geom_tufteboxplot(outlier.colour="black", outlier.shape = 22)+
```

```
  geom_boxplot()+
  theme_tufte() +
  theme(axis.title=element_blank()) +
  ylab('Scaled Values')+
  xlab('Variable')+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))+
  ggtitle('Distrobution of Values', subtitle = 'Y values scaled to fit a common axis')
df %>%
  select_if(is.factor) %>%
  gather() %>%
  ggplot(aes(x=value))+
  geom_bar()+
  facet_wrap(~key,scales='free_x')+
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
dfCont <- df %>% select(-TARGET_FLAG)
dfLog <- df %>% select(-TARGET_AMT)
mod1 <- lm(TARGET_AMT ~ ., data = dfCont)
#summary(mod1)
options(xtable.comment = FALSE)
summary(mod1)
mod2 <- lm(log(TARGET_AMT+1) ~ ., data = dfCont) # Note the `+1`
summary(mod2)

# Box Cox Method, univariate
summary(p1 <- powerTransform(I(TARGET_AMT+1) ~ ., dfCont))

bcTrans <- lm(I((TARGET_AMT+1)^(-0.4)) ~ ., dfCont)
summary(bcTrans)
logMod1 <- glm(TARGET_FLAG~., data = dfLog)
logMod1 %>% summary()
X <- dfLog %>% dplyr::select(-TARGET_FLAG)
X <- data.matrix(X)
#X <- as.matrix(X, ncol=12)
y <- as.factor(dfLog$TARGET_FLAG)
fit = glmnet(X, y, family = "binomial")
library(ROCR)
aucDF <- X
lasso.model = cv.glmnet(X, y, family = "binomial", type.measure = 'class')

aucDF$lasso.prob <- predict(lasso.model, type="response", newx = X, s = 'lambda.1se')
pred <- prediction(aucDF$lasso.prob, y)


cvfit = cv.glmnet(X, y, family = "binomial", type.measure = "class")
plot(cvfit)
coef(cvfit, s = "lambda.1se")
# theoDF <- dfLog %>% select(TARGET_FLAG, KIDSDRIV, AGE, OLDCLAIM, REVOKED, RED_CAR,)
# modTheo <- glm(TARGET_FLAG~., data = theoDF)
# summary(modTheo) %>% xtable()
regressionDiagnostic <- function(fit){
  ## https://www.statmethods.net/stats/rdiagnostics.html
  library(car) # Required
  print(outlierTest(fit))
  qqPlot(fit, main = 'QQ Plot')
  #av.Plots(fit)
  cutoff <- 4/((nrow(dfCont)-length(fit$coefficients)-2))
  plot(fit, which=4, cook.levels=cutoff)
  # Influence Plot
  influencePlot(fit,    id.method="identify", main="Influence Plot",
                sub="Circle size is proportial to Cook's Distance" )
  library(MASS)
  sresid <- studres(fit)
  hist(sresid, freq=FALSE,
       main="Distribution of Studentized Residuals")
  xfit<-seq(min(sresid),max(sresid),length=40)
  yfit<-dnorm(xfit)
  lines(xfit, yfit)
  print(ncvTest(fit))
  # plot studentized residuals vs. fitted values
  spreadLevelPlot(fit)
  print(durbinWatsonTest(fit))
}
regressionDiagnostic(mod1)
regressionDiagnostic(mod2)
regressionDiagnostic(bcTrans)
AUC <- function(df, mod, modelName){
  library(plotROC)
  library(pROC)
  prob = predict(mod,type = c("response"))
  dfLog$prob=prob
  p = ggplot(dfLog, aes(d = TARGET_FLAG, m = prob)) +
    geom_roc(n.cuts = 0) +
    ggtitle(paste('AUC Graph for', modelName)) +
    xlab("False Positive Fraction") +
    ylab('True Positive Fraction') +
    geom_abline(linetype = 'dashed') +
    theme_tufte()
```

```
  g <- roc(TARGET_FLAG ~ prob, data = dfLog)
  return(list(p, g$auc))
}
AUC(select(testDF, -TARGET_AMT), logMod1, 'Baisic GLM Model')

#AUC(select(testDF, -TARGET_AMT), modTheo, 'Theoretical Model')
#testX <- a
fittedGLMcv <- predict(cvfit, X, s = "lambda.1se", type = "class")

perf <- performance(pred,"tpr","fpr")
auc <-  performance(pred,"auc") # shows calculated AUC for model
auc <- auc@y.values

roc.data <- data.frame(fpr=unlist(perf@x.values),
                       tpr=unlist(perf@y.values))

ggplot(roc.data, aes(x=fpr, ymin=0, ymax=tpr)) +
    #geom_ribbon(alpha=0.2) +
    geom_line(aes(y=tpr)) +
    geom_abline(slope=1, intercept=0, linetype='dashed') +
    ggtitle("ROC Curve") +
    ylab('True Positive Rate') +
    xlab('False Positive Rate') + theme_tufte()
auc
#evalDF <- evalDF %>% select(-CONTAINS_NA)
predict(mod2,newdata = evalDF) %>% head()
predict(logMod1, newdata = evalDF) %>% head()
```