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.

Contents

1 Data Exploration **1**
1.1 The Data Frames . . . 1

2 Insert description of variables. **3**
2.1 Summary Statistics . . . 4
2.2 Descriptive Statistics . . . 4
2.3 Graphical EDA . . . 7

3 Data Preperation **9**
3.1 Transformed Skewed Variables . . . 9

4 Build Models **9**
4.1 Baisic Regression . . . 9
4.2 BoxCox . . . 12
4.3 Stepwise Selection on Baisic Model . . . 14
4.4 LASSO Regression . . . 14
4.5 Regular logostic . . . 14
4.6 Lasso with scaled variable . . . 14

5 Apendix **14**

1 Data Exploration

1.1 The Data Frames

Table 1: Sample of Values for the Training Set

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1
0	0	0	41	0	10	53401	No
1	4435	0	61	0	14	76875	No
1	3776	0	46	0	9	24202	No
1	6290	0	59	0	14	119537	No
0	0	0	30	3	13	18972	Yes
0	0	3	38	3	11	85100	No

Table 2: Sample of Values for the Training Set

HOME_VAL	MSTATUS	SEX	EDUCATION	JOB	TRAVTIME	CAR_USE	BLUEBOOK	TIF
0	z_No	z_F	Bachelors	z_Blue Collar	26	Commercial	14210	11
0	z_No	z_F	Masters	Lawyer	39	Private	17740	10
0	z_No	M	PhD	Lawyer	24	Private	17270	13
360697	Yes	z_F	PhD	Professional	26	Commercial	48380	4
119932	z_No	M	<High School	z_Blue Collar	29	Commercial	31460	6
256407	Yes	z_F	Masters	z_Blue Collar	34	Private	13810	1

Table 3: Sample of Values for the Training Set

CAR_TYPE	RED_CAR	OLDCLAIM	CLM_FREQ	REVOKED	MVR_PTS	CAR_AGE	URBANICITY
z_SUV	no	3189	2	No	1	1	Highly Urban/ Url
Sports Car	no	6587	3	No	8	1	Highly Urban/ Url
Van	yes	3475	1	No	4	17	Highly Urban/ Url
Panel Truck	no	0	0	Yes	1	20	Highly Urban/ Url
Panel Truck	yes	588	2	No	4	6	Highly Urban/ Url
Minivan	no	0	0	Yes	1	18	Highly Urban/ Url

```
## Observations: 4,534
## Variables: 25
## $ TARGET_FLAG <int> 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0,...
## $ TARGET_AMT  <dbl> 0.000, 0.000, 0.000, 2946.000, 2501.000, 0.000, 60...
## $ KIDSDRIV    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2,...
## $ AGE         <int> 60, 43, 35, 34, 34, 50, 53, 43, 55, 45, 39, 42, 34...
## $ HOMEKIDS    <int> 0, 0, 1, 1, 0, 0, 0, 0, 0, 0, 3, 0, 3, 2, 1, 0, 2,...
## $ YOJ         <int> 11, 11, 10, 12, 10, 7, 14, 5, 11, 0, 12, 11, 13, 1...
## $ INCOME      <dbl> 67349, 91449, 16039, 125301, 62978, 106952, 77100,...
## $ PARENT1     <fct> No, No, No, Yes, No, No, No, No, No, No, Yes, No, ...
## $ HOME_VAL    <dbl> 0, 257252, 124191, 0, 0, 0, 0, 209970, 180232, 106...
## $ MSTATUS     <fct> z_No, z_No, Yes, z_No, z_No, z_No, z_No, Yes, Yes,...
## $ SEX         <fct> M, M, z_F, z_F, z_F, M, z_F, z_F, M, z_F, z_F, M, ...
## $ EDUCATION   <fct> PhD, z_High School, z_High School, Bachelors, Bach...
## $ JOB         <fct> Professional, z_Blue Collar, Clerical, z_Blue Coll...
## $ TRAVTIME    <int> 14, 22, 5, 46, 34, 48, 15, 36, 25, 48, 43, 42, 27,...
## $ CAR_USE     <fct> Private, Commercial, Private, Commercial, Private,...
## $ BLUEBOOK    <dbl> 14230, 14940, 4010, 17430, 11200, 18510, 18300, 22...
## $ TIF         <int> 11, 1, 4, 1, 1, 7, 1, 7, 7, 1, 6, 6, 7, 4, 6, 10, ...
## $ CAR_TYPE    <fct> Minivan, Minivan, z_SUV, Sports Car, z_SUV, Van, S...
## $ RED_CAR     <fct> yes, yes, no, no, no, no, no, no, yes, no, no, no,...
## $ OLDCLAIM    <dbl> 4461, 0, 38690, 0, 0, 0, 0, 0, 5028, 0, 0, 0, 0, 0...
## $ CLM_FREQ    <int> 2, 0, 2, 0, 0, 0, 0, 0, 2, 0, 0, 0, 0, 0, 2, 0, 3,...
## $ REVOKED     <fct> No, No, No, No, No, No, No, No, Yes, No, No, No, N...
## $ MVR_PTS     <int> 3, 0, 3, 0, 0, 1, 0, 0, 3, 3, 0, 0, 0, 0, 0, 1, 0,...
## $ CAR_AGE     <int> 18, 1, 10, 7, 1, 17, 11, 1, 9, 5, 13, 16, 20, 7, 1...
## $ 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.

KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAL	MSTATUS	SEX	EDUCATION
0	43	0	14	24748	No	135512	Yes	M	<High School
0	48	0	13	38896	No	214165	Yes	M	<High School
0	52	0	8	174993	No	0	z_No	M	PhD
0	60	0	12	37940	No	182739	Yes	z_F	z_High School
0	32	3	10	22901	No	87839	Yes	M	z_High School
0	35	1	12	117579	No	296271	Yes	z_F	Bachelors

2 Insert description of variables.

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 th
CAR_AGE	Vehicle Age	Unknown effect on probability of collision, but probably effect th
CAR_TYPE	Type of Car	Unknown effect on probability of collision, but probably effect th
CAR_USE	Vehicle Use	Commercial vehicles are driven more, so might increase probabili
CLM_FREQ	# Claims (Past 5 Years)	The more claims you filed in the past, the more you are likely to
EDUCATION	Max Education Level	Unknown effect, but in theory more educated people tend to driv
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 cr
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 sugges
PARENT1	Single Parent	Unknown effect
RED_CAR	A red Car	Urban legend says that red cars (especially red sports cars) are m
REVOKED	License Revoked (Past 7 Years)	If your license was revoked in the past 7 years, you probably are
Sex	Time in Force	Urban legend says that women have less crashes then men. Is tha
TIF	Time in Force	People who have been customers for a long time are usually more
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

2.1 Summary Statistics

Table 6: 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 :0.2651	Mean : 1466	Mean :0.1665	Mean :44.72	Mean :0.7177	Mean :10.41
3rd Qu.:1.0000	3rd Qu.: 1102	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:13.00
Max. :1.0000	Max. :85524	Max. :4.0000	Max. :80.00	Max. :5.0000	Max. :23.00
NA	NA	NA	NA	NA	NA

Table 7: Summary Statistics

INCOME	PARENT1	HOME_VAL	MSTATUS	SEX	EDUCATION	JOB
Min. : 0	No :3926	Min. : 0	Yes :2683	M :1996	<High School : 718	z_Blue Collar:1114
1st Qu.: 26924	Yes: 608	1st Qu.: 0	z_No:1851	z_F:2538	Bachelors :1311	Clerical : 763
Median : 51624	NA	Median :158574	NA	NA	Masters : 804	Professional : 640
Mean : 58326	NA	Mean :149393	NA	NA	PhD : 314	Manager : 583
3rd Qu.: 81733	NA	3rd Qu.:232316	NA	NA	z_High School:1387	Lawyer : 519
Max. :367030	NA	Max. :885282	NA	NA	NA	Student : 395
NA	NA	NA	NA	NA	NA	(Other) : 520

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.

2.2 Descriptive Statistics

Table 8: Descriptive Statistics

	vars	n	mean	sd	median
TARGET_FLAG	1	4534	2.651081e-01	4.414394e-01	0
TARGET_AMT	2	4534	1.465843e+03	4.453326e+03	0
KIDSDRIV	3	4534	1.665196e-01	5.043988e-01	0
AGE	4	4534	4.472276e+01	8.686166e+00	45
HOMEKIDS	5	4534	7.176886e-01	1.112732e+00	0
YOJ	6	4534	1.041001e+01	4.158634e+00	11
INCOME	7	4534	5.832551e+04	4.404808e+04	51624
PARENT1*	8	4534	1.134098e+00	3.407951e-01	1
HOME_VAL	9	4534	1.493929e+05	1.239724e+05	158574
MSTATUS*	10	4534	1.408249e+00	4.915638e-01	1
SEX*	11	4534	1.559771e+00	4.964694e-01	2
EDUCATION*	12	4534	3.075210e+00	1.486713e+00	3
JOB*	13	4534	5.002426e+00	2.476582e+00	5
TRAVTIME	14	4534	3.377393e+01	1.592386e+01	33
CAR_USE*	15	4534	1.665858e+00	4.717417e-01	2
BLUEBOOK	16	4534	1.518569e+04	7.986098e+03	14070
TIF	17	4534	5.370754e+00	4.134133e+00	4

	vars	n	mean	sd	median
CAR_TYPE*	18	4534	3.554918e+00	2.000762e+00	3
RED_CAR*	19	4534	1.277018e+00	4.475749e-01	1
OLDCLAIM	20	4534	4.063686e+03	8.827165e+03	0
CLM_FREQ	21	4534	7.959859e-01	1.161585e+00	0
REVOKED*	22	4534	1.120203e+00	3.252345e-01	1
MVR_PTS	23	4534	1.691442e+00	2.175477e+00	1
CAR_AGE	24	4534	8.004632e+00	5.564949e+00	8
URBANICITY*	25	4534	1.213057e+00	4.095127e-01	1

Table 9: Descriptive Statistics

	trimmed	mad	min	max
TARGET_FLAG	2.064498e-01	0.0000	0	1.00
TARGET_AMT	5.982188e+02	0.0000	0	85523.65
KIDSDRIV	2.177510e-02	0.0000	0	4.00
AGE	4.471940e+01	8.8956	16	80.00
HOMEKIDS	4.939361e-01	0.0000	0	5.00
YOJ	1.099752e+01	2.9652	0	23.00
INCOME	5.405246e+04	39979.7916	0	367030.00
PARENT1*	1.042723e+00	0.0000	1	2.00
HOME_VAL	1.397377e+05	144228.8106	0	885282.00
MSTATUS*	1.385336e+00	0.0000	1	2.00
SEX*	1.574697e+00	0.0000	1	2.00
EDUCATION*	3.093991e+00	1.4826	1	5.00
JOB*	5.127894e+00	2.9652	1	8.00
TRAVTIME	3.325193e+01	16.3086	5	134.00
CAR_USE*	1.707277e+00	0.0000	1	2.00
BLUEBOOK	1.455607e+04	8080.1700	1500	65970.00
TIF	4.868523e+00	4.4478	1	25.00
CAR_TYPE*	3.568633e+00	2.9652	1	6.00
RED_CAR*	1.221334e+00	0.0000	1	2.00
OLDCLAIM	1.724477e+03	0.0000	0	53986.00
CLM_FREQ	5.843440e-01	0.0000	0	5.00
REVOKED*	1.025358e+00	0.0000	1	2.00
MVR_PTS	1.295755e+00	1.4826	0	13.00
CAR_AGE	7.602260e+00	5.9304	0	28.00
URBANICITY*	1.141400e+00	0.0000	1	2.00

Table 10: Descriptive Statistics

	range	skew	kurtosis	se
TARGET_FLAG	1.00	1.0639744	-0.8681498	0.0065559
TARGET_AMT	85523.65	8.4582529	104.6244189	66.1368787
KIDSDRIV	4.00	3.3856967	11.9227048	0.0074909
AGE	64.00	0.0038429	-0.0467281	0.1289993
HOMEKIDS	5.00	1.3436681	0.6575718	0.0165253
YOJ	23.00	-1.1820650	1.0360174	0.0617604
INCOME	367030.00	1.1211380	2.1733351	654.1633350
PARENT1*	1.00	2.1468700	2.6096266	0.0050612

	range	skew	kurtosis	se
HOME_VAL	885282.00	0.4207291	-0.1554888	1841.1292695
MSTATUS*	1.00	0.3732210	-1.8611164	0.0073003
SEX*	1.00	-0.2407296	-1.9424775	0.0073731
EDUCATION*	4.00	0.1404385	-1.4499284	0.0220793
JOB*	7.00	-0.3293866	-1.1747236	0.0367800
TRAVTIME	129.00	0.4372067	0.5055335	0.2364871
CAR_USE*	1.00	-0.7030177	-1.5060981	0.0070059
BLUEBOOK	64470.00	0.7725029	0.7093366	118.6025126
TIF	24.00	0.9036178	0.5294500	0.0613965
CAR_TYPE*	5.00	-0.0420538	-1.5407210	0.0297136
RED_CAR*	1.00	0.9961808	-1.0078460	0.0066470
OLDCLAIM	53986.00	3.0818959	9.5606550	131.0932927
CLM_FREQ	5.00	1.2142890	0.2772093	0.0172508
REVOKED*	1.00	2.3350123	3.4530443	0.0048301
MVR_PTS	13.00	1.4066158	1.6116292	0.0323083
CAR_AGE	28.00	0.3464418	-0.6259168	0.0826457
URBANICITY*	1.00	1.4010790	-0.0369858	0.0060817

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

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ	INCOME	PARENT1	HOME_VAI
0	0	0	0	0	0	0	0	0

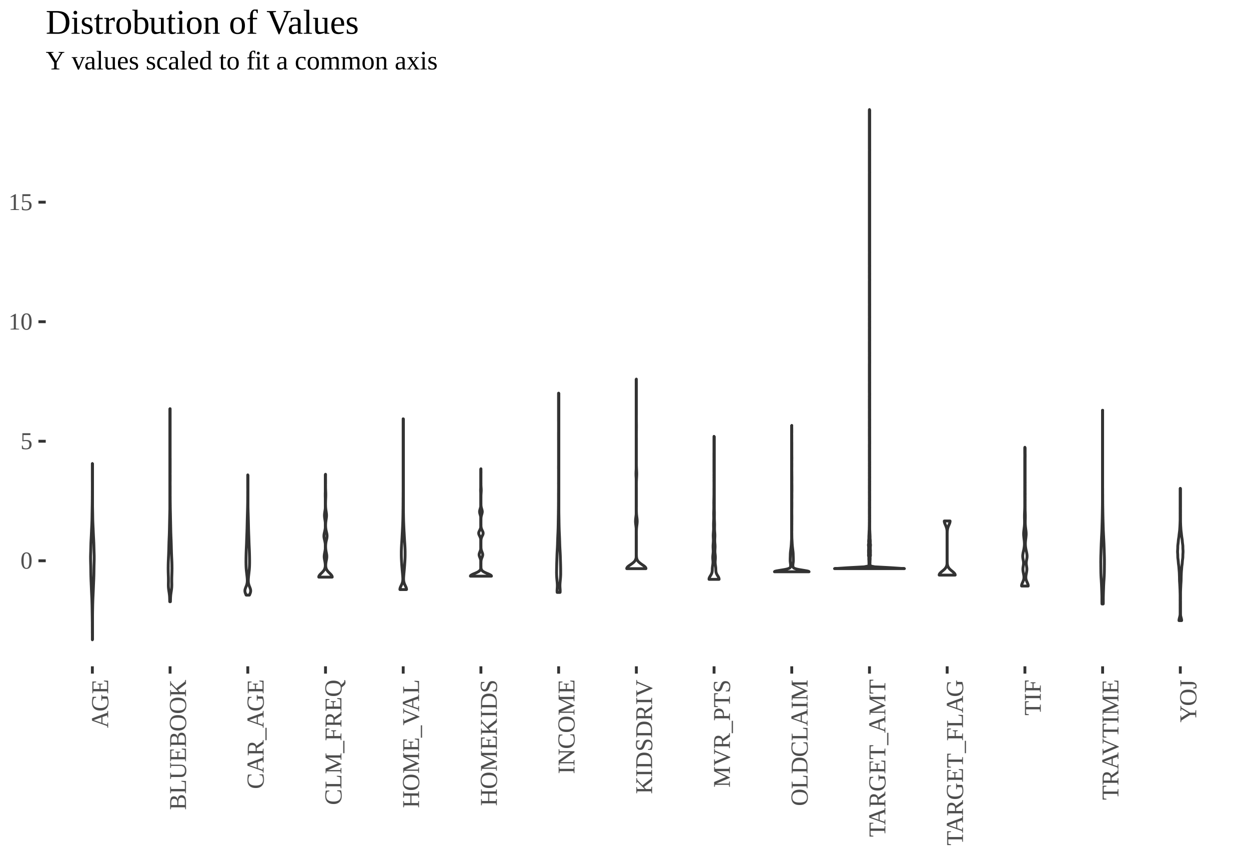
There are quite a few missing values accross several variables. However, compared to the size of the training set, around 6000, these numbers could be dropped if there is no correlation between the missing values and the response variables.

```
## Warning in cor(df$TARGET_FLAG, df$CONTAINS_NA): the standard deviation is
## zero
```

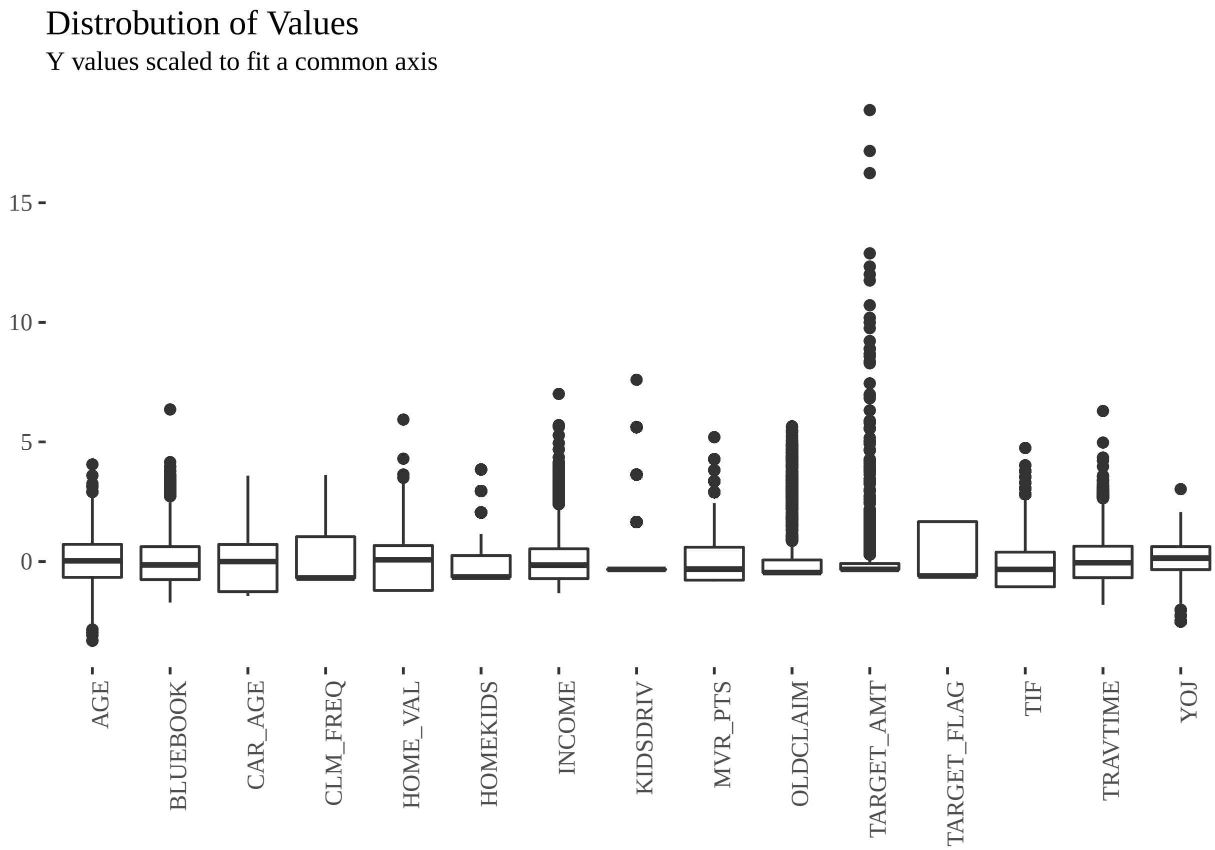
```
## Warning in cor(df$TARGET_AMT, df$CONTAINS_NA): the standard deviation is
## zero
```

The correlation between missing values and the 'Claim Filed' response is NA and NA 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.

2.3 Graphical EDA

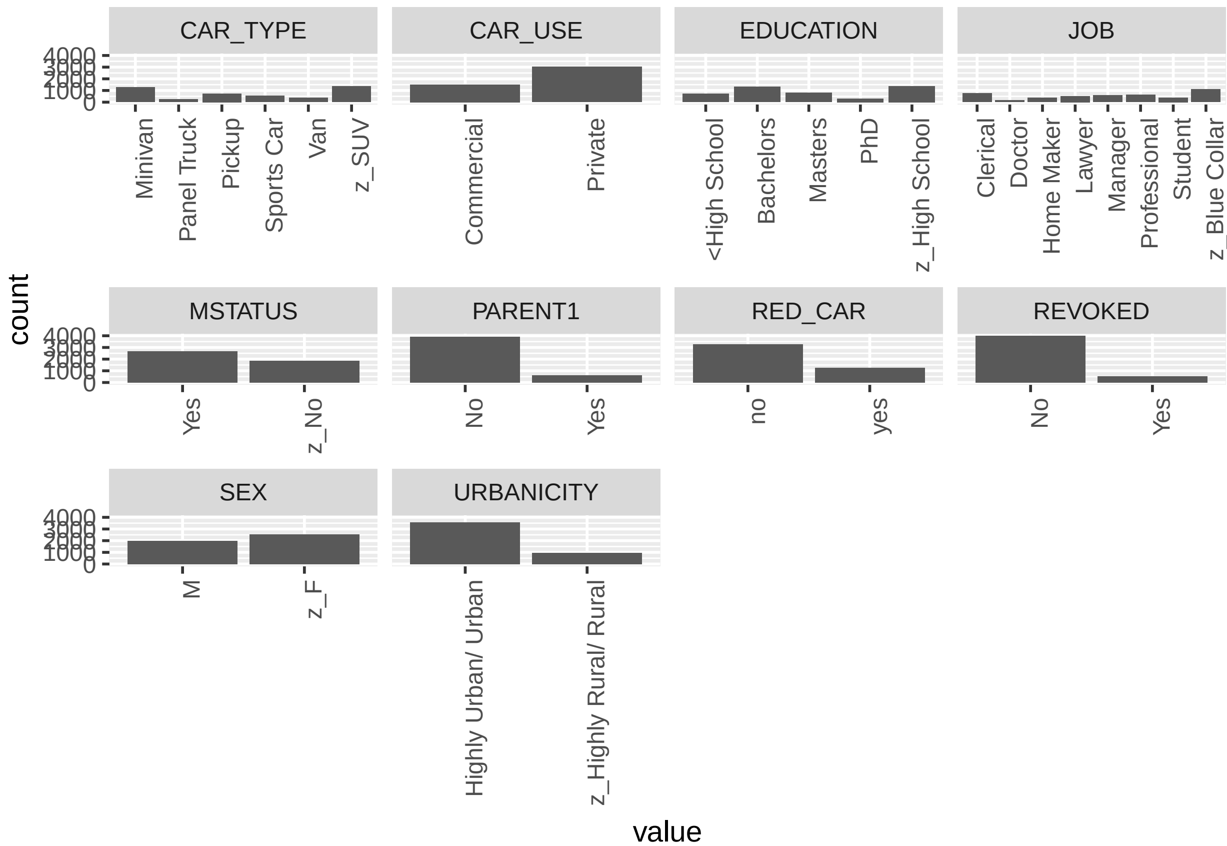


The distrobutions are generally skeded upwards, nothing suggests problems with the dataset. The only variable that is very skewed is `TARGET_AMT` and this makes sense because most are zero or low and some are very high.



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.

3 Data Preperation

3.1 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.

4 Build Models

This project will focus on automated variable selection. New techniques will be compared to the basic logistic regression.

4.1 Baisic Regression

Here I regressed all variables without transformation.

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

Residuals: Min 1Q Median 3Q Max -4977 -1650 -718 421 82800

Coefficients: (1 not defined because of singularities) Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.418e+03 5.968e+02 4.051 5.19e-05 ***KIDSDRIV 9.412e+01 1.446e+02 0.651 0.51509***
AGE -7.709e+00 8.871e+00 -0.869 0.38487
HOMEKIDS 5.228e+01 8.245e+01 0.634 0.52607
YOJ -1.122e+01 1.851e+01 -0.606 0.54456
INCOME -2.670e-03 2.544e-03 -1.049 0.29412
PARENT1Yes 4.297e+02 2.521e+02 1.705 0.08832 .
HOME_VAL -1.286e-03 8.029e-04 -1.602 0.10933
MSTATUSz_No 5.665e+02 1.879e+02 3.015 0.00258 *SEXz_F -6.748e+02 2.278e+02 -2.962 0.00307* ***EDUCATIONBachelors -2.147e+02 2.549e+02 -0.842 0.39960***
EDUCATIONMasters -9.384e+01 3.861e+02 -0.243 0.80799
EDUCATIONPhD 5.124e+02 4.768e+02 1.075 0.28258
EDUCATIONz_High School 8.603e+00 2.114e+02 0.041 0.96753
JOBDoctor -1.062e+03 5.609e+02 -1.893 0.05846 .
JOBHome Maker 1.116e+02 3.123e+02 0.357 0.72091
JOBLawyer -5.162e+01 3.794e+02 -0.136 0.89180
JOBManager -8.902e+02 2.936e+02 -3.032 0.00244 *JOBProfessional 1.366e+02 2.688e+02 0.508 0.61150*
JOBStudent -2.072e+02 3.010e+02 -0.688 0.49125
JOBz_Blue Collar 1.349e+02 2.382e+02 0.566 0.57110
TRAVTIME 1.334e+01 4.070e+00 3.277 0.00106 ***CAR_USEPrivate -8.360e+02 2.081e+02 -4.018 5.96e-05*** BLUEBOOK 1.422e-02 1.082e-02 1.314 0.18889
TIF -3.915e+01 1.547e+01 -2.530 0.01143 *
CAR_TYPEPanel Truck 1.886e+02 3.749e+02 0.503 0.61489
CAR_TYPEPickup 2.612e+02 2.133e+02 1.225 0.22069
CAR_TYPESports Car 1.247e+03 2.657e+02 4.694 2.76e-06 ***CAR_TYPEVan 4.020e+02 2.733e+02 1.471 0.14138***
CAR_TYPEz_SUV 9.683e+02 2.170e+02 4.462 8.32e-06 RED_CARyes -2.738e+02 1.938e+02 -1.413 0.15782
OLDCLAIM -7.220e-03 9.455e-03 -0.764 0.44512
CLM_FREQ 5.563e+01 7.014e+01 0.793 0.42774
REVOKEDYes 5.860e+02 2.219e+02 2.641 0.00829 ** MVR_PTS 1.480e+02 3.249e+01 4.556 5.36e-06
CAR_AGE -3.170e+01 1.637e+01 -1.936 0.05289 .
URBANICITYz_Highly Rural/ Rural -1.798e+03 1.743e+02 -10.317 < 2e-16 CONTAINS_NATRUE NA NA NA NA
— Signif. codes: 0 '' ***0.001*** '' *0.01* '' 0.05 '.' 0.1 '' 1

Residual standard error: 4284 on 4497 degrees of freedom Multiple R-squared: 0.08212, Adjusted R-squared: 0.07478 F-statistic: 11.18 on 36 and 4497 DF, p-value: < 2.2e-16

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2417.6207	596.8409	4.05	0.0001
KIDSDRIV	94.1160	144.5756	0.65	0.5151
AGE	-7.7091	8.8708	-0.87	0.3849
HOMEKIDS	52.2796	82.4511	0.63	0.5261
YOJ	-11.2182	18.5126	-0.61	0.5446
INCOME	-0.0027	0.0025	-1.05	0.2941
PARENT1Yes	429.6907	252.0662	1.70	0.0883
HOME_VAL	-0.0013	0.0008	-1.60	0.1093
MSTATUSz_No	566.5414	187.9129	3.01	0.0026
SEXz_F	-674.7755	227.8053	-2.96	0.0031
EDUCATIONBachelors	-214.7004	254.8637	-0.84	0.3996
EDUCATIONMasters	-93.8447	386.1469	-0.24	0.8080
EDUCATIONPhD	512.3776	476.7734	1.07	0.2826
EDUCATIONz_High School	8.6035	211.3522	0.04	0.9675
JOBDoctor	-1061.5456	560.8613	-1.89	0.0585
JOBHome Maker	111.5825	312.3292	0.36	0.7209
JOBLawyer	-51.6180	379.4324	-0.14	0.8918
JOBManager	-890.2411	293.6022	-3.03	0.0024
JOBProfessional	136.5535	268.8189	0.51	0.6115
JOBStudent	-207.2371	301.0495	-0.69	0.4912
JOBz_Blue Collar	134.9214	238.1778	0.57	0.5711
TRAVTIME	13.3393	4.0700	3.28	0.0011
CAR_USEPrivate	-835.9945	208.0600	-4.02	0.0001
BLUEBOOK	0.0142	0.0108	1.31	0.1889
TIF	-39.1495	15.4726	-2.53	0.0114
CAR_TYPEPanel Truck	188.6265	374.8962	0.50	0.6149
CAR_TYPEPickup	261.2268	213.2684	1.22	0.2207
CAR_TYPESports Car	1247.2655	265.7158	4.69	0.0000
CAR_TYPEVan	401.9563	273.2695	1.47	0.1414
CAR_TYPEz_SUV	968.2973	217.0155	4.46	0.0000
RED_CARyes	-273.7794	193.8002	-1.41	0.1578
OLDCLAIM	-0.0072	0.0095	-0.76	0.4451
CLM_FREQ	55.6294	70.1382	0.79	0.4277
REVOKEDYes	586.0490	221.8963	2.64	0.0083
MVR_PTS	148.0291	32.4928	4.56	0.0000
CAR_AGE	-31.7033	16.3729	-1.94	0.0529
URBANICITYz_Highly Rural/ Rural	-1797.8027	174.2644	-10.32	0.0000

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

```
##
## Call:
## lm(formula = log(TARGET_AMT + 1) ~ ., data = dfCont)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.4280 -2.3433 -0.8535  2.1207 10.1690
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      3.473e+00  4.465e-01   7.777 9.14e-15 ***
## KIDSDRIV                         2.876e-01  1.082e-01   2.659 0.007870 **
## AGE                             -4.307e-03  6.637e-03  -0.649 0.516410
## HOMEKIDS                         6.679e-02  6.169e-02   1.083 0.279006
```

```
## YOJ                              -1.587e-02  1.385e-02  -1.145 0.252076
## INCOME                           -2.411e-06  1.904e-06  -1.267 0.205370
## PARENT1Yes                        6.430e-01  1.886e-01   3.410 0.000656 ***
## HOME_VAL                         -1.510e-06  6.007e-07  -2.514 0.011979 *
## MSTATUSz_No                       5.355e-01  1.406e-01   3.809 0.000141 ***
## SEXz_F                           -3.710e-01  1.704e-01  -2.177 0.029549 *
## EDUCATIONBachelors               -4.175e-01  1.907e-01  -2.190 0.028601 *
## EDUCATIONMasters                 -4.686e-01  2.889e-01  -1.622 0.104890
## EDUCATIONPhD                      1.155e-01  3.567e-01   0.324 0.746101
## EDUCATIONz_High School            7.024e-02  1.581e-01   0.444 0.656918
## JOBDoctor                        -1.192e+00  4.196e-01  -2.840 0.004525 **
## JOBHome Maker                    -2.072e-01  2.337e-01  -0.887 0.375338
## JOBLawyer                        -2.681e-01  2.839e-01  -0.945 0.344960
## JOBManager                       -1.355e+00  2.197e-01  -6.168 7.52e-10 ***
## JOBProfessional                  -4.143e-01  2.011e-01  -2.060 0.039475 *
## JOBStudent                       -2.447e-01  2.252e-01  -1.086 0.277370
## JOBz_Blue Collar                 -1.030e-01  1.782e-01  -0.578 0.563310
## TRAVTIME                          2.016e-02  3.045e-03   6.620 4.02e-11 ***
## CAR_USEPrivate                   -1.011e+00  1.557e-01  -6.496 9.17e-11 ***
## BLUEBOOK                         -2.505e-05  8.096e-06  -3.094 0.001985 **
## TIF                              -6.705e-02  1.158e-02  -5.792 7.42e-09 ***
## CAR_TYPEPanel Truck               6.530e-01  2.805e-01   2.328 0.019947 *
## CAR_TYPEPickup                    5.611e-01  1.596e-01   3.516 0.000442 ***
## CAR_TYPESports Car                1.319e+00  1.988e-01   6.635 3.63e-11 ***
## CAR_TYPEVan                       4.311e-01  2.044e-01   2.109 0.035018 *
## CAR_TYPEz_SUV                     9.438e-01  1.624e-01   5.813 6.56e-09 ***
## RED_CARyes                       -3.043e-01  1.450e-01  -2.099 0.035879 *
## OLDCLAIM                         -1.865e-05  7.074e-06  -2.637 0.008395 **
## CLM_FREQ                          2.491e-01  5.247e-02   4.747 2.13e-06 ***
## REVOKEDYes                        1.126e+00  1.660e-01   6.782 1.33e-11 ***
## MVR_PTS                           2.127e-01  2.431e-02   8.749  < 2e-16 ***
## CAR_AGE                          -1.041e-02  1.225e-02  -0.849 0.395656
## URBANICITYz_Highly Rural/ Rural -2.474e+00  1.304e-01 -18.979  < 2e-16 ***
## CONTAINS_NATRUE                          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.205 on 4497 degrees of freedom
## Multiple R-squared:  0.2446, Adjusted R-squared:  0.2386
## F-statistic: 40.46 on 36 and 4497 DF,  p-value: < 2.2e-16
```

Here we have the output of all provided variables without any transformation.

4.2 BoxCox

```
## bcPower Transformation to Normality
##    Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## Y1   -0.3967        -0.4      -0.4121      -0.3814
##
## Likelihood ratio tests about transformation parameters
##                            LRT df pval
## LR test, lambda = (0)  3316.157  1    0
## LR test, lambda = (1) 48706.172  1    0
```

```
##
## Call:
## lm(formula = I((TARGET_AMT + 1)^(-0.4)) ~ ., data = dfCont)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08639 -0.26531  0.09827  0.27263  0.86171
##
## Coefficients: (1 not defined because of singularities)
##                                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      5.968e-01  5.155e-02  11.576  < 2e-16 ***
## KIDSDRIV                        -3.571e-02  1.249e-02  -2.860 0.004258 **
## AGE                              5.066e-04  7.662e-04   0.661 0.508552
## HOMEKIDS                        -7.071e-03  7.122e-03  -0.993 0.320796
## YOJ                              1.602e-03  1.599e-03   1.002 0.316381
## INCOME                           2.538e-07  2.198e-07   1.155 0.248127
## PARENT1Yes                      -7.418e-02  2.177e-02  -3.407 0.000662 ***
## HOME_VAL                         1.721e-07  6.935e-08   2.482 0.013103 *
## MSTATUSz_No                     -6.082e-02  1.623e-02  -3.747 0.000181 ***
## SEXz_F                           3.920e-02  1.968e-02   1.992 0.046413 *
## EDUCATIONBachelors               4.915e-02  2.201e-02   2.233 0.025617 *
## EDUCATIONMasters                 5.737e-02  3.335e-02   1.720 0.085507 .
## EDUCATIONPhD                    -8.585e-03  4.118e-02  -0.208 0.834870
## EDUCATIONz_High School          -7.310e-03  1.826e-02  -0.400 0.688853
## JOBDoctor                        1.364e-01  4.844e-02   2.816 0.004876 **
## JOBHome Maker                    2.380e-02  2.698e-02   0.882 0.377689
## JOBLawyer                        3.070e-02  3.277e-02   0.937 0.348978
## JOBManager                       1.573e-01  2.536e-02   6.201 6.11e-10 ***
## JOBProfessional                  5.148e-02  2.322e-02   2.217 0.026669 *
## JOBStudent                       2.446e-02  2.600e-02   0.941 0.346873
## JOBz_Blue Collar                 1.348e-02  2.057e-02   0.655 0.512441
## TRAVTIME                        -2.347e-03  3.515e-04  -6.676 2.75e-11 ***
## CAR_USEPrivate                   1.171e-01  1.797e-02   6.516 8.02e-11 ***
## BLUEBOOK                         3.185e-06  9.347e-07   3.407 0.000663 ***
## TIF                              7.782e-03  1.336e-03   5.823 6.18e-09 ***
## CAR_TYPEPanel Truck             -7.613e-02  3.238e-02  -2.351 0.018766 *
## CAR_TYPEPickup                  -6.582e-02  1.842e-02  -3.573 0.000356 ***
## CAR_TYPESports Car              -1.498e-01  2.295e-02  -6.529 7.37e-11 ***
## CAR_TYPEVan                     -4.925e-02  2.360e-02  -2.087 0.036979 *
## CAR_TYPEz_SUV                   -1.057e-01  1.874e-02  -5.638 1.82e-08 ***
## RED_CARyes                       3.632e-02  1.674e-02   2.170 0.030071 *
## OLDCLAIM                         2.249e-06  8.167e-07   2.754 0.005906 **
## CLM_FREQ                        -3.087e-02  6.058e-03  -5.096 3.62e-07 ***
## REVOKEDYes                      -1.318e-01  1.917e-02  -6.877 6.96e-12 ***
## MVR_PTS                         -2.419e-02  2.807e-03  -8.618  < 2e-16 ***
## CAR_AGE                          1.068e-03  1.414e-03   0.755 0.450216
## URBANICITYz_Highly Rural/ Rural  2.860e-01  1.505e-02  19.003  < 2e-16 ***
## CONTAINS_NATRUE                         NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.37 on 4497 degrees of freedom
## Multiple R-squared:  0.2464, Adjusted R-squared:  0.2404
## F-statistic: 40.85 on 36 and 4497 DF,  p-value: < 2.2e-16
```

```
## bcPower Transformations to Multinormality
##       Est Power Rounded Pwr Wald Lwr bnd Wald Upr Bnd
## len      0.1451        0.00      -0.2733       0.5636
## adt      0.2396        0.33       0.0255       0.4536
## trks    -0.7336        0.00      -1.9408       0.4735
## sigs1   -0.2959       -0.50      -0.5511      -0.0408
##
## Likelihood ratio tests about transformation parameters
##                                   LRT df       pval
## LR test, lambda = (0 0 0 0)  13.1339  4 0.01063972
## LR test, lambda = (1 1 1 1) 140.5853  4 0.00000000
```

4.3 Stepwise Selection on Baisic Model

If we do a step wise selection to find the variables that limit the scope but still provide excellent performance we get:

4.4 LASSO Regression

Looking at lasso logistic regression might give us a better model selection and coefficient values. Below is the results.

4.5 Regular logostic

This models coefficients deviate significantly from a normal `glm` model that excludes the one variable dropped. This is because LASSO penalizes large coefficients. For example, `glm` model excluding `rm` is:

This is interesting because we can see how different the coefficients are even though it has the same variables.

4.6 Lasso with scaled variable

5 Apendix

```r
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_knit$set(root.dir = "/Users/kailukowiak/DATA621/Assignments")
# 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)
```

```r
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)
LabledDF <- LabledDF[complete.cases(LabledDF), ]
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[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)
evalDF %>% sample_n(6) %>% 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')
```

```r
######################
dis <- describe(df)
dis[, 1:5] %>% kable(caption = 'Descriptive Statistics')
dis[, 6:9] %>% kable(caption = 'Descriptive Statistics')
dis[, 10:13] %>% kable(caption = 'Descriptive Statistics')
map(df, ~sum(is.na(.))) %>% t() %>% 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 %>%
  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)
xtable(summary(mod1))
mod2 <- lm(log(TARGET_AMT+1) ~ ., data = dfCont) # Note the `+1`
```

```r
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)
#summary(powerTransform(cbind(len, adt, trks, sigs1) ~ 1, Highway1))
summary(a3 <- powerTransform(cbind(len, adt, trks, sigs1) ~ htype, Highway1))
```