Homework #4: Insurance Claim Prediction

Data 621 Business Analytics and Data Mining

Aadi Kalloo, Nathan Lim, Asher Meyers, Daniel Smilowitz, Logan Thomson

Due July 10, 2016

Contents

1 Data Exploration

The dataset of interest contains information about customers of an auto insurance company. The dataset has 8161 rows (each representing a customer) and 25 variables. There are 23 predictor variables and 2 response variables: `TARGET_FLAG`, a binary categorical variable representing whether each customer has been in an accident; and `TARGET_AMT`, a numerical variable indicating the cost of a crash that a customer was in. The class of variables read in from the dataset is presented below:

	Class	Levels
TARGET_FLAG	integer	-
TARGET_AMT	numeric	-
KIDSDRIV	integer	-
AGE	integer	-
HOMEKIDS	integer	-
YOJ	integer	-
INCOME	factor	6613
PARENT1	factor	2
HOME_VAL	factor	5107
MSTATUS	factor	2
SEX	factor	2
EDUCATION	factor	5
JOB	factor	9
TRAVTIME	integer	-
CAR_USE	factor	2
BLUEBOOK	factor	2789
TIF	integer	-
CAR_TYPE	factor	6
RED_CAR	factor	2
OLDCLAIM	factor	2857
CLM_FREQ	integer	-
REVOKED	factor	2
MVR_PTS	integer	-
CAR_AGE	integer	-
URBANICITY	factor	2

The very high number of levels for four of the variables (`INCOME`, `HOME_VAL`, `BLUEBOOK`, and `OLDCLAIM`) indicates that these variables are not in fact factors; investigation of the dataset indicates that these are dollar values interpreted as strings due to the presence of dollar signs and commas. The numerical values are extracted for these variables.

Additionally, there are 7 variables with only two levels. These are recast as binary variables as follows:

- `PARENT1`, `MSTATUS`, `RED_CAR`, and `REVOKED`: using 1 to indicate Yes
- `SEX`: using 1 to indicate Male
- `CAR_USE`: using 1 to indicate Commercial
- `URBANICITY`: using 1 to indicate Highly Urban/ Urban

Finally, there are three categorical variables – factors with more than two levels. Dummy variables are created for each of these, as follows:

- `EDUCATION`: using High School as the base case
- `CAR_TYPE`: using Minivan as the base case
- `JOB`: using the blank value as the base case

A summary of each variable is presented below:

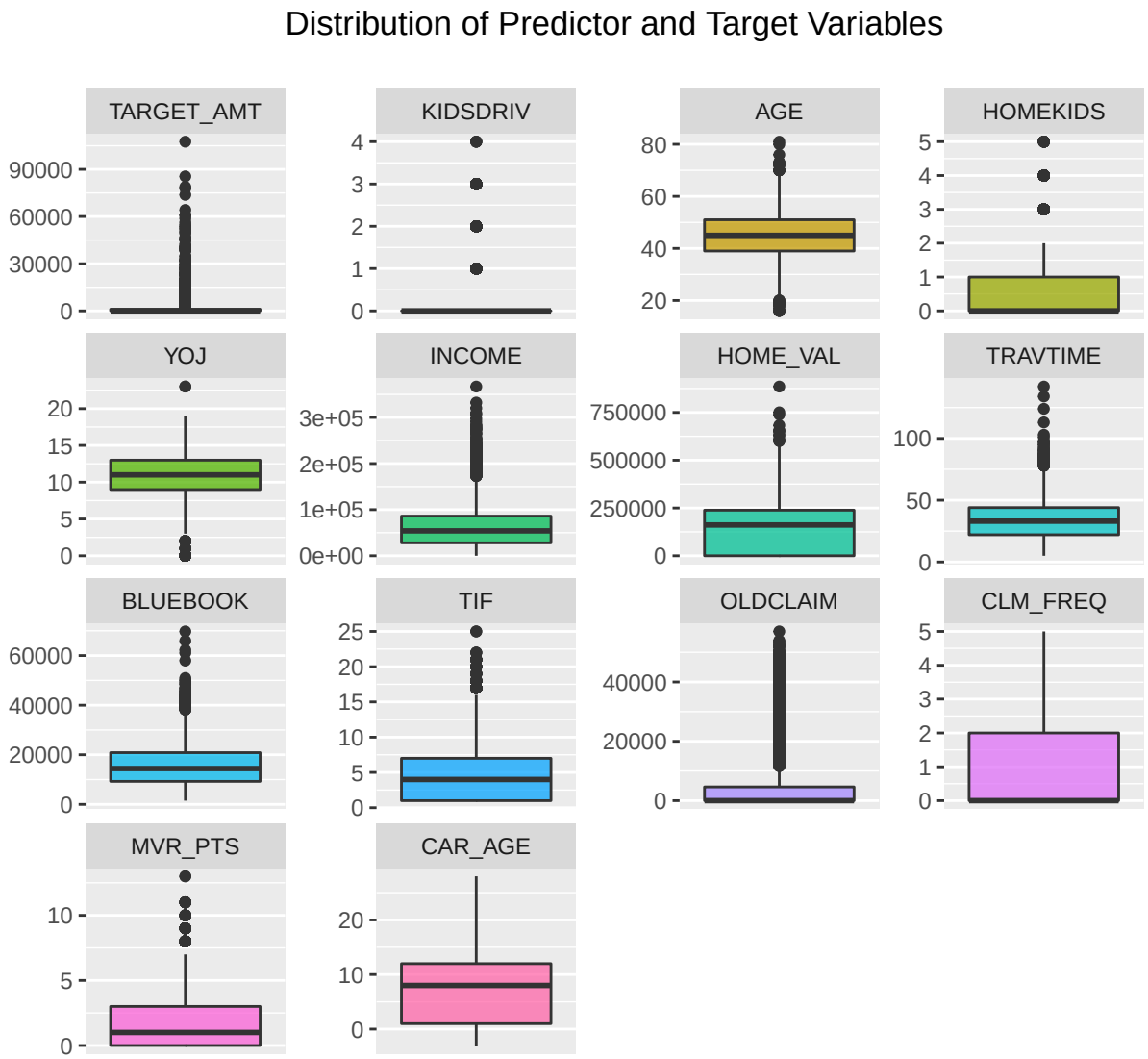
	MEAN	MEDIAN	IQR	SKEW	r_{FLAG}	r_{AMT}	NAs
TARGET_FLAG	0.26	0	1	1.07	1	0.54	0
TARGET_AMT	1504	0	1036	8.71	0.54	1	0
KIDSDRIV	0.17	0	0	3.35	0.09	0.05	0
AGE	44.79	45	12	-0.03	-0.11	-0.05	6
HOMEKIDS	0.72	0	1	1.34	0.11	0.06	0
YOJ	10.5	11	4	-1.2	-0.07	-0.02	454
INCOME	61898	54028	57889	1.19	-0.14	-0.06	445
PARENT1	0.13	0	0	2.17	0.16	0.1	0
HOME_VAL	154867	161160	238724	0.49	-0.18	-0.09	464
MSTATUS	0.6	1	1	-0.41	-0.13	-0.1	0
SEX	0.46	0	1	0.14	-0.02	0.01	0
TRAVTIME	33.49	33	22	0.45	0.05	0.03	0
CAR_USE	0.37	0	1	0.53	0.14	0.1	0
BLUEBOOK	15710	14440	11570	0.79	-0.11	0	0
TIF	5.35	4	6	0.89	-0.08	-0.04	0
RED_CAR	0.29	0	1	0.92	-0.02	0	0
OLDCLAIM	4037	0	4636	3.12	0.14	0.08	0
CLM_FREQ	0.8	0	2	1.21	0.22	0.12	0
REVOKED	0.12	0	0	2.3	0.15	0.06	0
MVR_PTS	1.7	1	3	1.35	0.23	0.14	0
CAR_AGE	8.33	8	11	0.28	-0.11	-0.06	510
URBANICITY	0.8	1	0	-1.46	0.22	0.12	0
HSDropout	0.15	0	0	1.99	0.06	0.04	0
Bachelors	0.27	0	1	1.01	-0.05	-0.02	0
Masters	0.2	0	0	1.48	-0.09	-0.05	0
PhD	0.09	0	0	2.88	-0.06	-0.02	0
Panel_Truck	0.08	0	0	3.03	0	0.04	0
Pickup	0.17	0	0	1.75	0.05	0.02	0
Sports_Car	0.11	0	0	2.47	0.06	0.03	0
Van	0.09	0	0	2.82	0	0.01	0
SUV	0.28	0	1	0.97	0.05	0.01	0
Professional	0.14	0	0	2.11	-0.04	0	0
Blue_Collar	0.14	0	0	2.11	-0.04	0	0
Clerical	0.16	0	0	1.9	0.04	0	0
Doctor	0.03	0	0	5.49	-0.05	-0.03	0
Lawyer	0.1	0	0	2.62	-0.06	-0.03	0
Manager	0.12	0	0	2.32	-0.12	-0.07	0
Home_Maker	0.08	0	0	3.13	0.01	0	0
Student	0.09	0	0	2.92	0.07	0.02	0

From the table above, it is clear that there are four variables with missing values, with the proportion of values missing ranging from less than $< 0.1\%$ to roughly 6.2%; these missing values will need to either be imputed or excluded from the dataset before modeling. The variables exhibit varying levels of skewness, with a few extreme values.

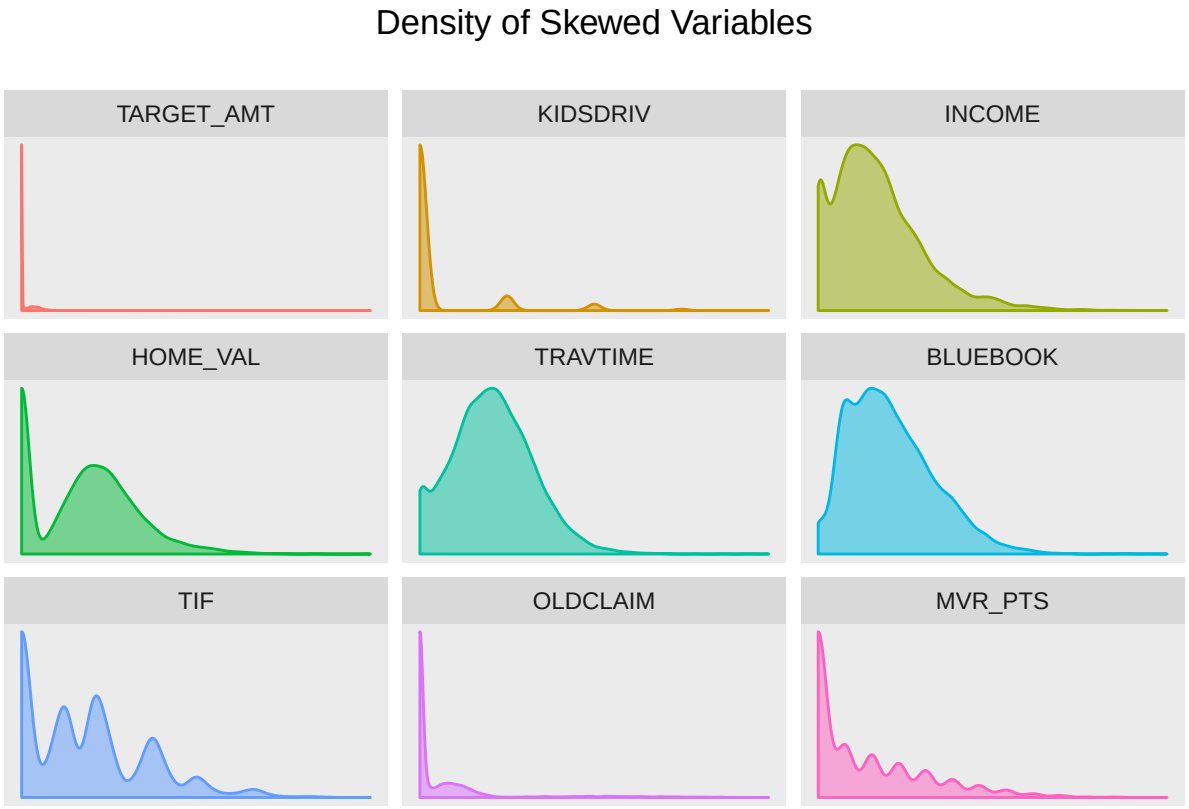
The large number of binary variables in the dataset makes graphical visualization of the distribution of all variables not particularly useful. The proportion of binary variables having a value of 0 or 1 is presented in the table below:

	0	1
TARGET_FLAG	0.74	0.26
PARENT1	0.87	0.13
MSTATUS	0.4	0.6
SEX	0.54	0.46
CAR_USE	0.63	0.37
RED_CAR	0.71	0.29
REVOKED	0.88	0.12
URBANICITY	0.2	0.8
HSDropout	0.85	0.15
Bachelors	0.73	0.27
Masters	0.8	0.2
PhD	0.91	0.09
Panel_Truck	0.92	0.08
Pickup	0.83	0.17
Sports_Car	0.89	0.11
Van	0.91	0.09
SUV	0.72	0.28
Professional	0.86	0.14
Blue_Collar	0.86	0.14
Clerical	0.84	0.16
Doctor	0.97	0.03
Lawyer	0.9	0.1
Manager	0.88	0.12
Home_Maker	0.92	0.08
Student	0.91	0.09

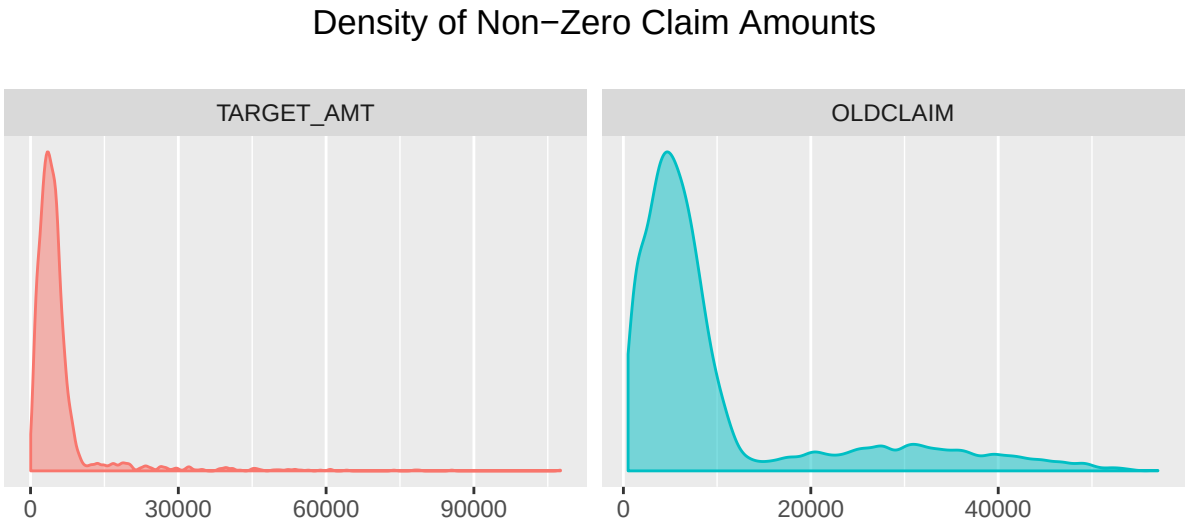
The remaining variables are vizualized below in boxplots:



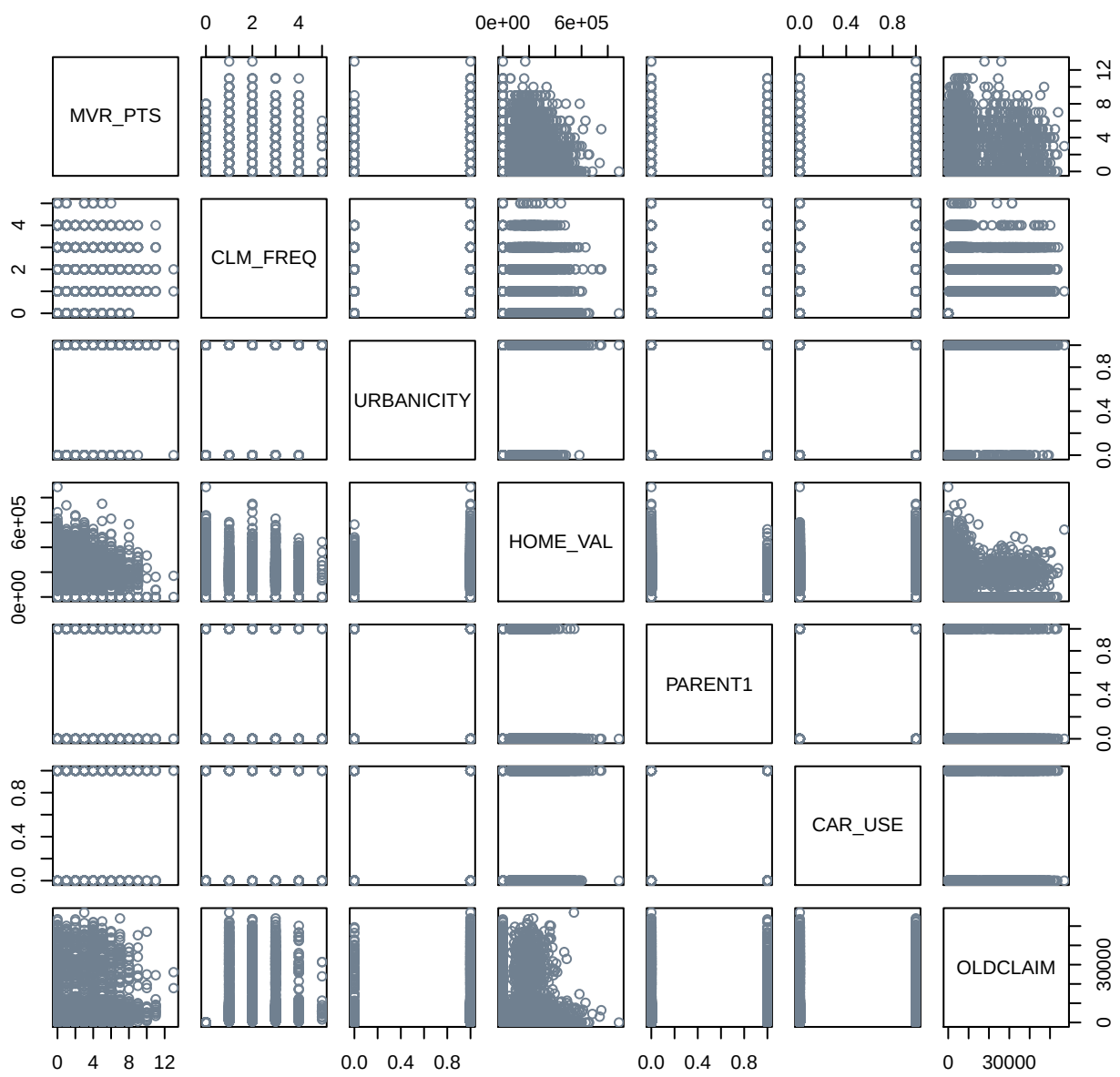
The boxplots illustrate the high skewness of the distributions of the predictors `KIDSDRIV`, `INCOME`, `HOME_VAL`, `TRAVTIME`, `BLUEBOOK`, `TIF`, `OLDCLAIM` and `MVR_PTS`. The target `TARGET_AMT` is also highly skewed – this makes sense, as this value is 0 for any customers without claims. Density plots of these variables are presented below:



Since `TARGET_AMT` and `OLDCLAIM` have such high concentrations at values of zero, separate plots for these two variables are created with values of zero removed:

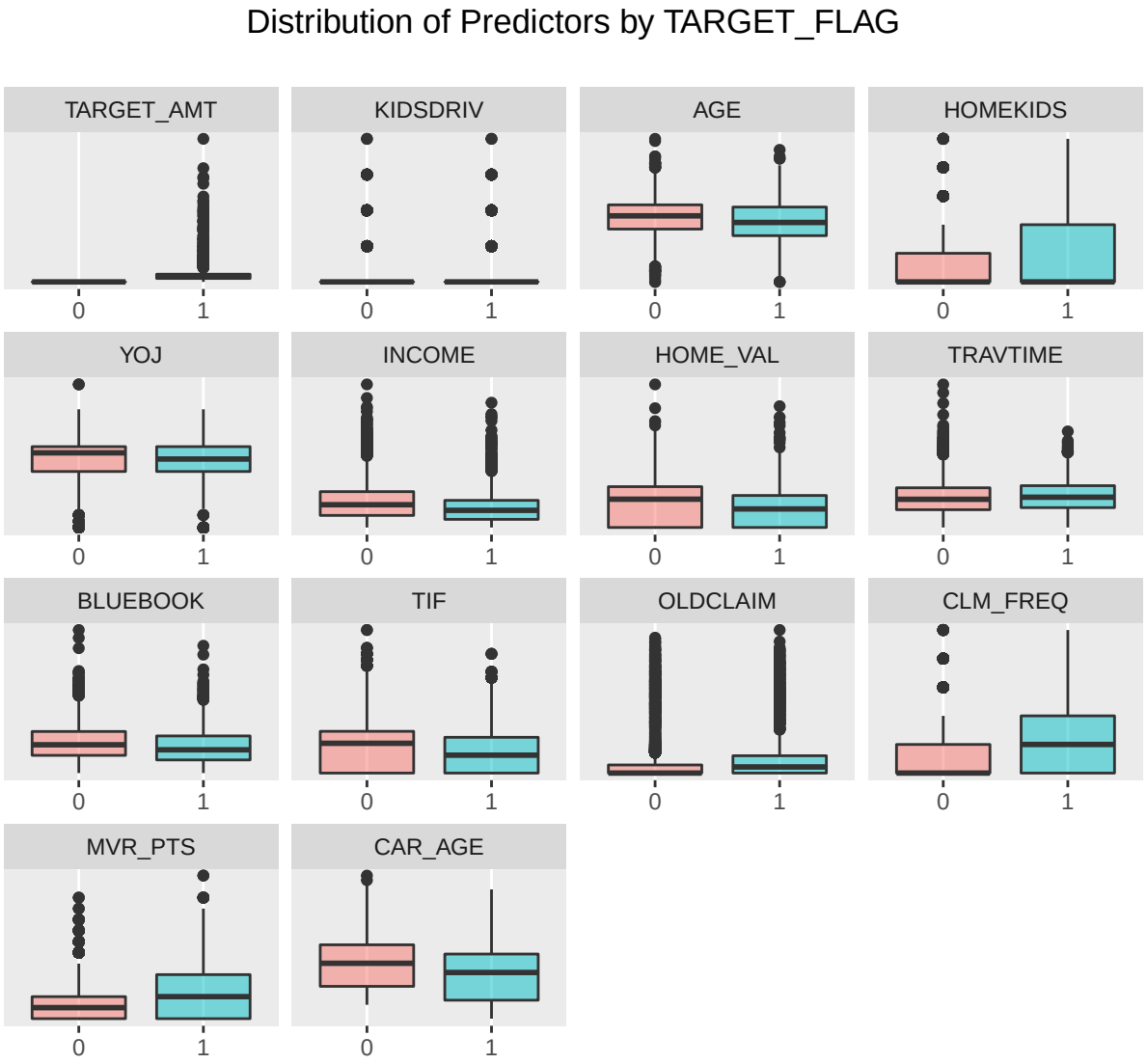


The 8 predictors with the highest correlation to `TARGET_FLAG` and the 8 predictors with the highest correlation to `TARGET_AMT` share 7 predictors. The correlation between these variables is investigated and plotted below:



There appears to be evidence of possible multicollinearity between `HOME_VAL` and `OLDCLAIM`.

Finally, boxplots are also prepared for non-binary variables split by `TARGET_FLAG`:



Interestingly, there is not an immediately visible difference in median value based on `TARGET_FLAG` values, while the range of predictor values for each flag value differs noticeably.

2 Data Preparation

3 Model Creation

3.1 Multiple Linear Regression

3.2 Binary Logistic Regression

4 Model Selection & Prediction

Appendix A: Index-wise Results from Predictive Model

Appendix B: R Code