

DATA 621 - Homework #4

Claudio, Mauricio

2022-04-22

// Overview

In this homework assignment, you will explore, analyze and model a data set containing approximately 8,000 records representing a customer at an auto insurance company. Each record has two response variables. The first response variable, TARGET_FLAG, is a 1 or a 0. A 1 means that the person was in a car crash. A zero means that the person was not in a car crash. The second response variable is TARGET_AMT. This value is zero if the person did not crash their car. But if they did crash their car, this number will be a value greater than zero.

Your objective is to build multiple linear regression and binary logistic regression models on the training data to predict the probability that a person will crash their car and also the amount of money it will cost if the person does crash their car. You can only use the variables given to you (or variables that you derive from the variables provided). Below is a short description of the variables of interest in the data set:

VARIABLE NAME	DEFINITION	THEORETICAL EFFECT
INDEX	Identification Variable (do not use)	None
TARGET_FLAG	Was Car in a crash? 1=YES 0=NO	None
TARGET_AMT	If car was in a crash, what was the cost	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	Marital 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	Gender	Urban legend says that women have less crashes than 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
URBANICITY	Home/Work Area	Unknown
YOJ	Years on Job	People who stay at a job for a long time are usually more safe

// Data Exploration and Preparation

The raw dataset consists of 8,161 instances and 26 attributes. We note the following gaps in the data:

- Missing values in predictor variables AGE(0.1%), YOJ(5.6%), INCOME(5.5%), HOME_VAL (5.7%) and CAR_AGE (6.2%)
- Non-sensical negative values in predictor variable CAR_AGE

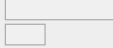

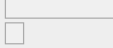
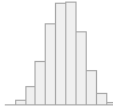
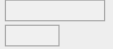
- Numerical predictor variables INCOME, HOME_VAL, BLUEBOOK and OLDCLAIM are erroneously encoded as character type

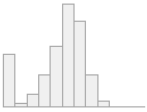
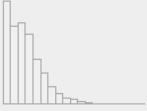
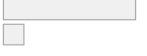
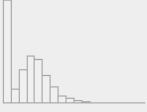




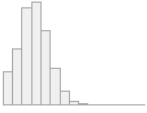
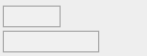
To prepare the data for model building, we effect the following transformations:

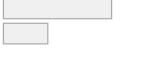

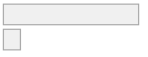
- Convert INCOME, HOME_VAL, BLUEBOOK and OLDCLAIM to numeric
- Convert missing values in YOJ, INCOME, HOME_VAL and CAR_AGE to zero value
- Eliminate negative values in CAR_AGE by taking the absolute value
- Impute mean to missing values in AGE
- Convert categorical variables to factors

In addition we simplify categorical predictor variables EDUCATION, JOB, CAR_TYPE and CLM_FREQ by reducing their levels to fewer but more significant and meaningful levels. The latter two factor variables are mutated to variables sportscar and claims.

We do not test the data for colinearity, linearity or outlier/leverage points because our aim is prediction rather than inference and those tests are more appropriate in the context of a particular regression model or set of models. Likewise, we wish to avoid the potential for overfitting and diminished predictive performance that variable elimination due to colinearity or observation deletion could effect. Linearity will be addressed later in the context of regression model variable selection and transformation. The transformed dataset, ready for model building, is summarized below.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	TARGET_FLAG [integer]	Min : 0 Mean : 0.3 Max : 1	0 : 6008 (73.6%) 1 : 2153 (26.4%)		8161 (100.0%)	0 (0.0%)
2	TARGET_AMT [numeric]	Mean (sd) : 1504.3 (4704) min ≤ med ≤ max: 0 ≤ 0 ≤ 107586.1 IQR (CV) : 1036 (3.1)	1949 distinct values		8161 (100.0%)	0 (0.0%)
3	KIDSDRIV [factor]	1. NO 2. YES	7180 (88.0%) 981 (12.0%)		8161 (100.0%)	0 (0.0%)
4	AGE [numeric]	Mean (sd) : 44.8 (8.6) min ≤ med ≤ max: 16 ≤ 45 ≤ 81 IQR (CV) : 12 (0.2)	61 distinct values		8161 (100.0%)	0 (0.0%)
5	HOMEKIDS [factor]	1. NO 2. YES	5289 (64.8%) 2872 (35.2%)		8161 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
6	YOJ [numeric]	Mean (sd) : 9.9 (4.6) min ≤ med ≤ max: 0 ≤ 11 ≤ 23 IQR (CV) : 5 (0.5)	21 distinct values		8161 (100.0%)	0 (0.0%)
7	INCOME [numeric]	Mean (sd) : 58522.9 (48345.5) min ≤ med ≤ max: 0 ≤ 51116 ≤ 367030 IQR (CV) : 60147 (0.8)	6612 distinct values		8161 (100.0%)	0 (0.0%)
8	PARENT1 [factor]	1. NO 2. YES	7084 (86.8%) 1077 (13.2%)		8161 (100.0%)	0 (0.0%)
9	HOME_VAL [numeric]	Mean (sd) : 146062.2 (130426.7) min ≤ med ≤ max: 0 ≤ 151957 ≤ 885282 IQR (CV) : 233352 (0.9)	5106 distinct values		8161 (100.0%)	0 (0.0%)
10	MSTATUS [factor]	1. Married 2. Unmarried	4894 (60.0%) 3267 (40.0%)		8161 (100.0%)	0 (0.0%)
11	SEX [factor]	1. Female 2. Male	4375 (53.6%) 3786 (46.4%)		8161 (100.0%)	0 (0.0%)
12	EDUCATION [factor]	1. HS or lower 2. University	3533 (43.3%) 4628 (56.7%)		8161 (100.0%)	0 (0.0%)
13	JOB [factor]	1. White Collar 2. Blue Collar 3. Other	4457 (54.6%) 1825 (22.4%) 1879 (23.0%)		8161 (100.0%)	0 (0.0%)
14	TRAVTIME [integer]	Mean (sd) : 33.5 (15.9) min ≤ med ≤ max: 5 ≤ 33 ≤ 142 IQR (CV) : 22 (0.5)	97 distinct values		8161 (100.0%)	0 (0.0%)
15	CAR_USE [factor]	1. Commercial 2. Private	3029 (37.1%) 5132 (62.9%)		8161 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
16	BLUEBOOK [numeric]	Mean (sd) : 15709.9 (8419.7) min ≤ med ≤ max: 1500 ≤ 14440 ≤ 69740 IQR (CV) : 11570 (0.5)	2789 distinct values		8161 (100.0%)	0 (0.0%)
17	TIF [integer]	Mean (sd) : 5.4 (4.1) min ≤ med ≤ max: 1 ≤ 4 ≤ 25 IQR (CV) : 6 (0.8)	23 distinct values		8161 (100.0%)	0 (0.0%)
18	RED_CAR [factor]	1. NO 2. YES	5783 (70.9%) 2378 (29.1%)		8161 (100.0%)	0 (0.0%)
19	OLDCLAIM [numeric]	Mean (sd) : 4037.1 (8777.1) min ≤ med ≤ max: 0 ≤ 0 ≤ 57037 IQR (CV) : 4636 (2.2)	2857 distinct values		8161 (100.0%)	0 (0.0%)
20	REVOKED [factor]	1. NO 2. YES	7161 (87.7%) 1000 (12.3%)		8161 (100.0%)	0 (0.0%)
21	MVR_PTS [integer]	Mean (sd) : 1.7 (2.1) min ≤ med ≤ max: 0 ≤ 1 ≤ 13 IQR (CV) : 3 (1.3)	13 distinct values		8161 (100.0%)	0 (0.0%)
22	CAR_AGE [numeric]	Mean (sd) : 7.8 (5.9) min ≤ med ≤ max: 0 ≤ 8 ≤ 28 IQR (CV) : 11 (0.8)	29 distinct values		8161 (100.0%)	0 (0.0%)
23	URBANITY [factor]	1. Rural 2. Urban	1669 (20.5%) 6492 (79.5%)		8161 (100.0%)	0 (0.0%)
24	sportscar [factor]	1. NO 2. YES	7254 (88.9%) 907 (11.1%)		8161 (100.0%)	0 (0.0%)
25	claims [factor]	1. No 2. Yes	5009 (61.4%) 3152 (38.6%)		8161 (100.0%)	0 (0.0%)

// Binary Logistic Regression

// Model Building and Selection

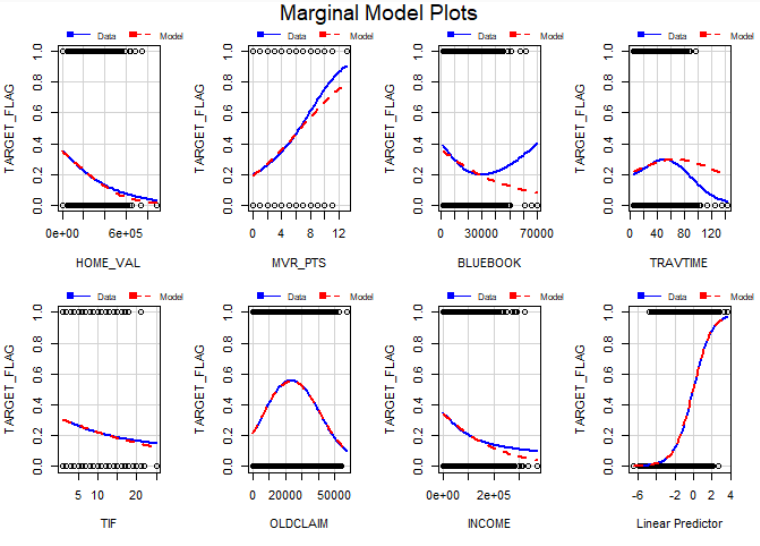
We build and select three binary logistic regression models to predict the outcome variable TARGET_FLAG. To do so, we perform backward and forward variable selection at 95% confidence level (i.e. p-value < 0.05) on all dataset predictor variables. Model A (AIC 7,451) is the base model obtained from the convergence of backward and forward selection. Model B (AIC 7,431) is the base model with transformed predictor variables. Model C (AIC 7,417) is Model B with two predictors removed and interaction terms added.

The three models are summarized with model coefficients shown as probabilities and their marginal model plots are displayed below.

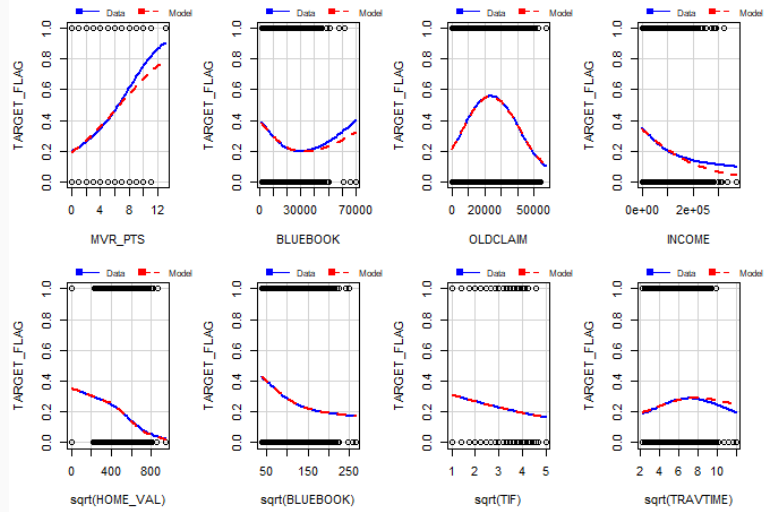
Model A			Model B		Model C	
Predictors	Probabilities	p	Probabilities	p	Probabilities	p
(Intercept)	0.07	5e-55	0.14	2e-08	0.10	7e-20
BLUEBOOK	0.50	2e-11	0.50	3e-02		
URBANICITY: Rural	Reference		Reference		Reference	
URBANICITY: Urban	0.90	8e-90	0.91	2e-90	0.90	1e-89
CAR_USE: Commercial	Reference		Reference		Reference	
CAR_USE: Private	0.28	1e-48	0.28	9e-47	0.28	3e-47
EDUCATION: HS or lower	Reference		Reference		Reference	
EDUCATION: University	0.37	2e-15	0.37	1e-15	0.37	7e-15
HOME_VAL	0.50	7e-04				
REVOKED: NO	Reference		Reference		Reference	
REVOKED: YES	0.73	5e-27	0.73	4e-27	0.72	5e-30
I(BLUEBOOK * INCOME)					0.50	1e-03
I(sqrt(OLDCLAIM * MVR_PTS))					0.50	3e-12
INCOME	0.50	9e-07	0.50	7e-09	0.50	7e-09
claims: No	Reference		Reference		Reference	
claims: Yes	0.65	2e-16	0.66	1e-16	0.67	3e-20
KIDSDRIV: NO	Reference		Reference		Reference	
KIDSDRIV: YES	0.63	6e-09	0.64	5e-09	0.63	7e-09
MVR_PTS	0.53	2e-14	0.53	2e-14	0.55	1e-25
OLDCLAIM	0.50	2e-07	0.50	2e-07		
sportscar: NO	Reference		Reference		Reference	

sportscar: YES	0.61	5e-06	0.60	2e-05	0.60	3e-05
MSTATUS: Married	Reference		Reference		Reference	
MSTATUS: Unmarried	0.66	5e-23	0.66	2e-21	0.66	2e-21
SEX: Female	Reference		Reference		Reference	
SEX: Male	0.45	4e-03	0.45	3e-03	0.45	5e-03
sqrt(BLUEBOOK)			0.50	3e-04	0.50	1e-14
sqrt(HOME_VAL)			0.50	6e-04	0.50	2e-03
sqrt(TIF)			0.44	3e-14	0.44	1e-14
sqrt(TRAVTIME)			0.54	1e-16	0.54	2e-16
TIF	0.49	1e-13				
TRAVTIME	0.50	8e-16				
HOMEKIDS: NO	Reference		Reference		Reference	
HOMEKIDS: YES	0.58	7e-07	0.58	1e-06	0.58	2e-06
Observations	8161		8161		8161	
R ² Tjur	0.237		0.240		0.244	

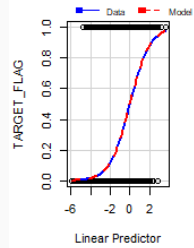
| Model A



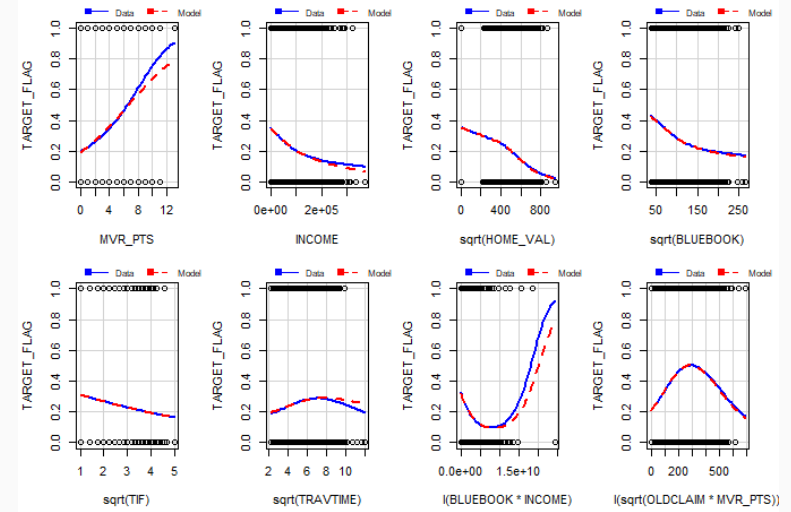
Model B



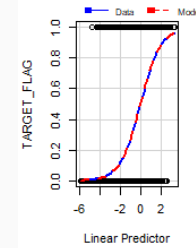
Marginal Model Plots



Model C



Marginal Model Plots

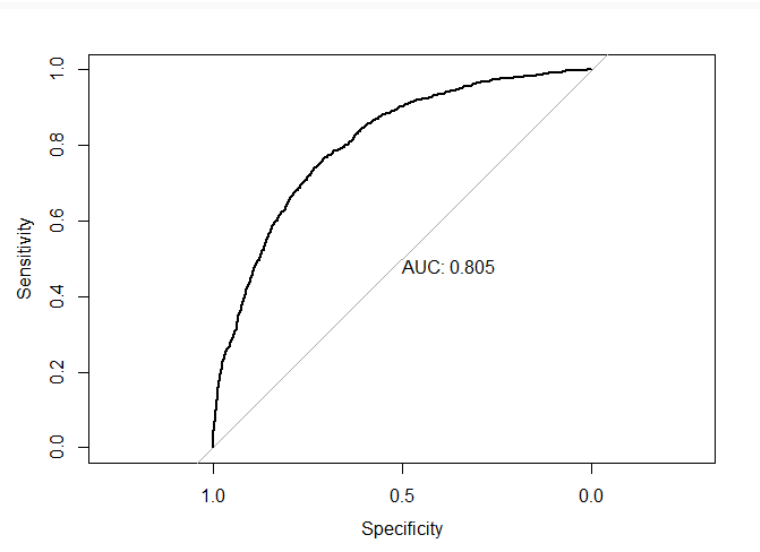
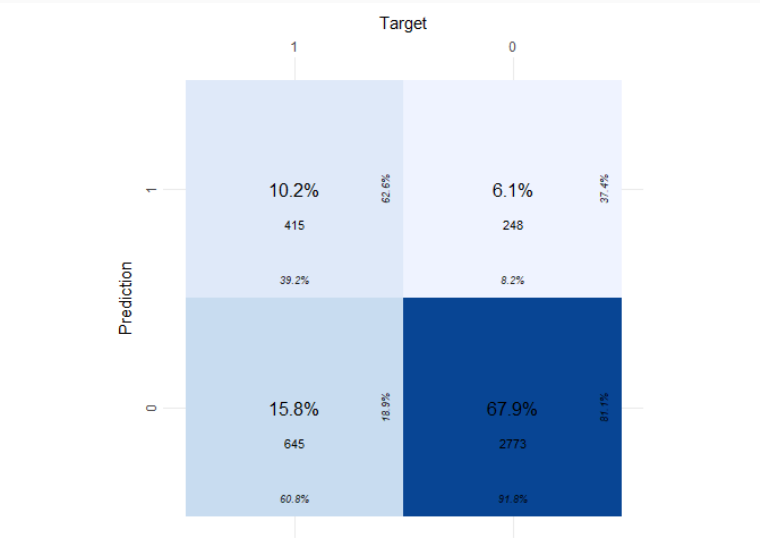


We select Models B and C for testing based on their superior AIC and model fit. The predictive performance of the two models is tested by Monte Carlo cross-validation with a 50% train / 50% test split and 1,000 iterations. **Based on the results of this testing, we select Model C based of its marginally better Accuracy and AUC.**

The confusion matrix, performance metrics and AUC curve for each model are shown below:

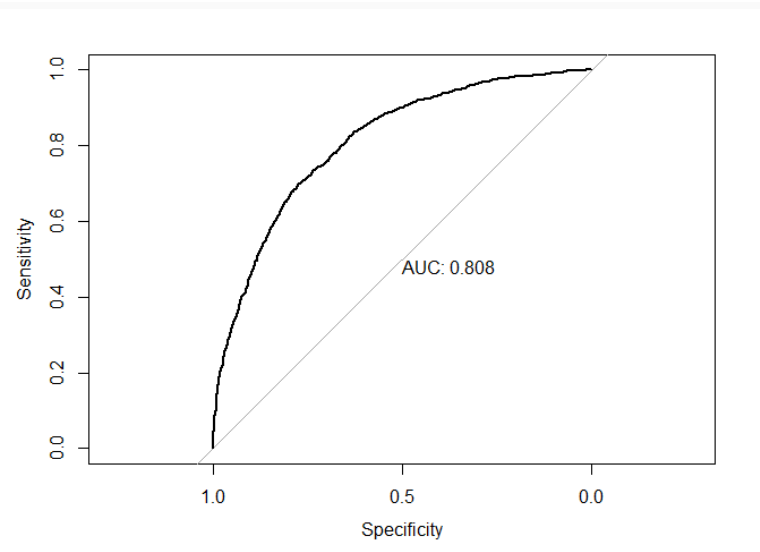
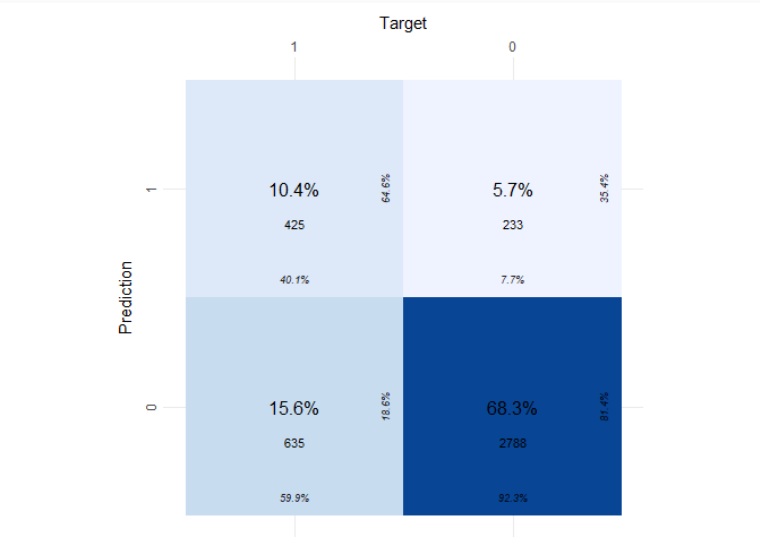
| Model B

Accuracy	Sensitivity	Specificity	Error	F1	Precision	validations
0.782	0.4	0.918	0.218	0.491	0.637	1000



| Model C

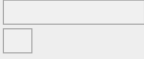
Accuracy	Sensitivity	Specificity	Error	F1	Precision	validations
0.785	0.407	0.92	0.215	0.499	0.646	1000



// Prediction

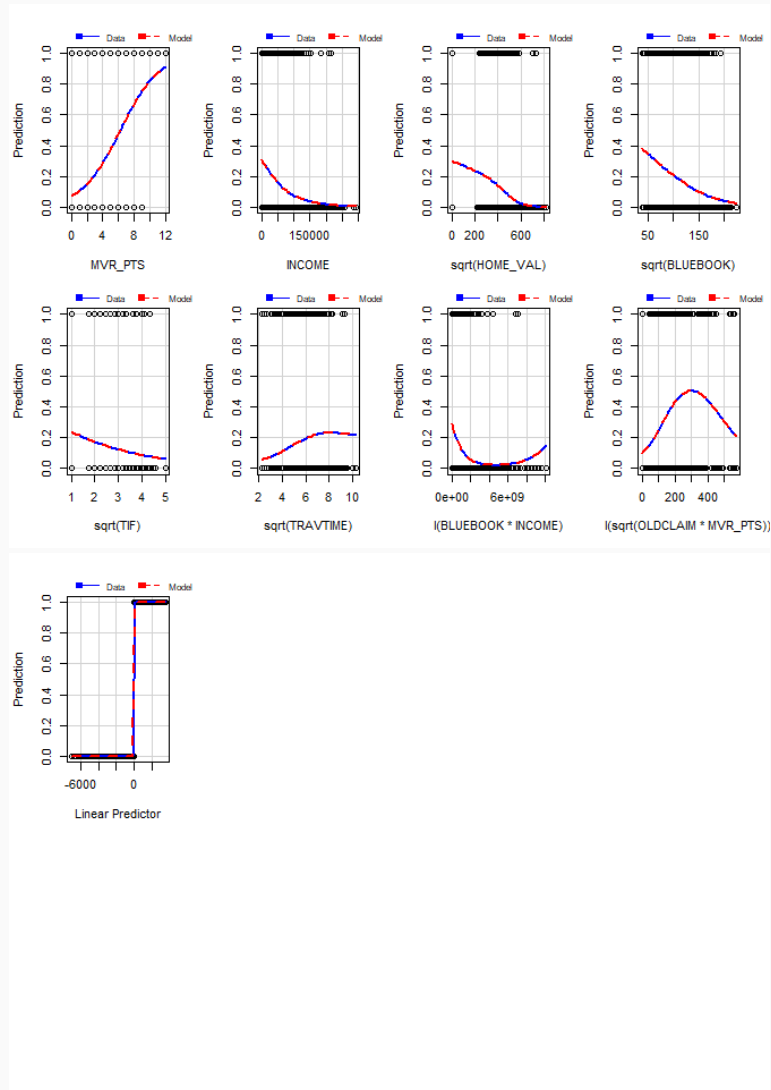
We transform the evaluation dataset in the same manner that we transformed the evaluation dataset earlier and predict the response based on our selected model, Model C. The predicted probabilities and predictions for the 2,141 observations in the evaluation data set are summarized below. A .csv file of the predictions is available for [download](#).

| Evaluation dataset: predictions summary

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid
Prediction [numeric]	Min : 0 Mean : 0.2 Max : 1	0 : 1774 (82.9%) 1 : 367 (17.1%)		2141 (100.0%)

Generated by [summarytools](#) 1.0.0 (R version 4.1.0)
2022-04-22

| Evaluation dataset: model summary and predicted probabilities



// Linear Regression

// Model Building and Selection

The data for the prediction of claim amount is that for which a crash has taken place. This corresponds to all dataset instances for which TARGET_FLAG equals one. This data is summarized below.

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	TARGET_AMT [numeric]	Mean (sd) : 5702.2 (7743.2) min ≤ med ≤ max: 30.3 ≤ 4104 ≤ 107586.1 IQR (CV) : 3177.2 (1.4)	1948 distinct values		2153 (100.0%)	0 (0.0%)
2	KIDSDRIV [factor]	1. NO 2. YES	1773 (82.4%) 380 (17.6%)		2153 (100.0%)	0 (0.0%)
3	AGE [numeric]	Mean (sd) : 43.3 (9.6) min ≤ med ≤ max: 16 ≤ 43 ≤ 76 IQR (CV) : 13 (0.2)	57 distinct values		2153 (100.0%)	0 (0.0%)
4	HOMEKIDS [factor]	1. NO 2. YES	1173 (54.5%) 980 (45.5%)		2153 (100.0%)	0 (0.0%)
5	YOJ [numeric]	Mean (sd) : 9.4 (5) min ≤ med ≤ max: 0 ≤ 11 ≤ 19 IQR (CV) : 5 (0.5)	19 distinct values		2153 (100.0%)	0 (0.0%)
6	INCOME [numeric]	Mean (sd) : 48054 (43140.9) min ≤ med ≤ max: 0 ≤ 41367 ≤ 320127 IQR (CV) : 53522 (0.9)	1734 distinct values		2153 (100.0%)	0 (0.0%)
7	PARENT1 [factor]	1. NO 2. YES	1677 (77.9%) 476 (22.1%)		2153 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
8	HOME_VAL [numeric]	Mean (sd) : 108779.1 (117811.3) min ≤ med ≤ max: 0 ≤ 99468 ≤ 750455 IQR (CV) : 193855 (1.1)	1155 distinct values		2153 (100.0%)	0 (0.0%)
9	MSTATUS [factor]	1. Married 2. Unmarried	1053 (48.9%) 1100 (51.1%)		2153 (100.0%)	0 (0.0%)
10	SEX [factor]	1. Female 2. Male	1192 (55.4%) 961 (44.6%)		2153 (100.0%)	0 (0.0%)
11	EDUCATION [factor]	1. HS or lower 2. University	1178 (54.7%) 975 (45.3%)		2153 (100.0%)	0 (0.0%)
12	JOB [factor]	1. White Collar 2. Blue Collar 3. Other	937 (43.5%) 634 (29.4%) 582 (27.0%)		2153 (100.0%)	0 (0.0%)
13	TRAVTIME [integer]	Mean (sd) : 34.8 (15.2) min ≤ med ≤ max: 5 ≤ 34 ≤ 97 IQR (CV) : 21 (0.4)	79 distinct values		2153 (100.0%)	0 (0.0%)
14	CAR_USE [factor]	1. Commercial 2. Private	1047 (48.6%) 1106 (51.4%)		2153 (100.0%)	0 (0.0%)
15	BLUEBOOK [numeric]	Mean (sd) : 14255.9 (8299.8) min ≤ med ≤ max: 1500 ≤ 12600 ≤ 62240 IQR (CV) : 11450 (0.6)	1398 distinct values		2153 (100.0%)	0 (0.0%)
16	TIF [integer]	Mean (sd) : 4.8 (3.9) min ≤ med ≤ max: 1 ≤ 4 ≤ 21 IQR (CV) : 6 (0.8)	19 distinct values		2153 (100.0%)	0 (0.0%)
17	RED_CAR [factor]	1. NO 2. YES	1537 (71.4%) 616 (28.6%)		2153 (100.0%)	0 (0.0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
18	OLDCLAIM [numeric]	Mean (sd) : 6061.5 (10071.1) min ≤ med ≤ max: 0 ≤ 2448 ≤ 57037 IQR (CV) : 6906 (1.7)	1205 distinct values		2153 (100.0%)	0 (0.0%)
19	REVOKED [factor]	1. NO 2. YES	1710 (79.4%) 443 (20.6%)		2153 (100.0%)	0 (0.0%)
20	MVR_PTS [integer]	Mean (sd) : 2.5 (2.6) min ≤ med ≤ max: 0 ≤ 2 ≤ 13 IQR (CV) : 4 (1)	13 distinct values		2153 (100.0%)	0 (0.0%)
21	CAR_AGE [numeric]	Mean (sd) : 6.9 (5.7) min ≤ med ≤ max: 0 ≤ 7 ≤ 25 IQR (CV) : 10 (0.8)	26 distinct values		2153 (100.0%)	0 (0.0%)
22	URBANITY [factor]	1. Rural 2. Urban	115 (5.3%) 2038 (94.7%)		2153 (100.0%)	0 (0.0%)
23	sportscar [factor]	1. NO 2. YES	1849 (85.9%) 304 (14.1%)		2153 (100.0%)	0 (0.0%)
24	claims [factor]	1. No 2. Yes	898 (41.7%) 1255 (58.3%)		2153 (100.0%)	0 (0.0%)

Generated by [summarytools](#) 1.0.0 (R version 4.1.0)
2022-04-22

We build two linear regression models at 95% confidence through backward selection and AIC minimization. Model Log Normal is Ordinary Least Squares with log transformed response. Model Gamma is a log-link General Linear Model (GLM) appropriate for right skew response variables. We also attempted to fit a log-link Inverse Gauss GLM model, but faced repeated difficulties with model convergence both with base R and the Tweedie family parameter with the *statmod* package. This suggests that an Inverse Gauss model was perhaps too extreme for the data at hand.

The Log Normal and Gamma GLM models are summarized below.

Model Log Normal

$\log(TARGET_AMT) = \log(BLUEBOOK) + MSTATUS + \sqrt{MVR_PTS}$
AIC 40815

Model GLM Gamma

$TARGET_AMT = \log(BLUEBOOK)$
AIC 41282

Log Normal			GLM Gamma	
Predictors	Estimates	p	Estimates	p
(Intercept)	6.689	5.1e-138	446.439	6.3e-57
log(BLUEBOOK)	0.160	1.2e-09	1.310	3.7e-11
MSTATUS: Married	Reference		Reference	
MSTATUS: Unmarried	0.073	3.6e-02		
sqrt(MVR_PTS)	0.040	2.3e-02		
Observations	2153		2153	
R ² / R ² adjusted	0.021 / 0.020		0.057	

We test the two models with 50% train / 50% test Monte Carlo cross-validation and 1000 iterations and calculate the Root Mean Square Error (RMSE) for each model. We use the RMSE because it is sensitive to the large errors that which wish to minimize in our predictions later on. For replicability we set the seed to `set.seed(i*10)` where i is the iteration. This method assures both that each iteration results in a different random sample test/train split and replicability.

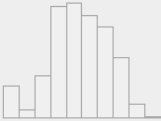
Based on the superior or lower RMSE value of Model GLM Gamma relative to Model Log Normal, we select Model GLM Gamma as our model for prediction.

RMSE: Model Log Norm	RMSE: Model Gamma	cross-validations
7884	7668	1000

// Prediction

Using Model GLM Gamma, we predict the claim amount for the 367 observations for which an accident was predicted in the evaluation data set. The predictions are summarized below and in a .csv file available for [download](#).

| Evaluation dataset: predictions summary

Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid
prediction [numeric]	Mean (sd) : 5413.9 (940.6) min ≤ med ≤ max: 3217.2 ≤ 5444.1 ≤ 7663.2 IQR (CV) : 1315.1 (0.2)	305 distinct values		367 (100.0%)