Modeling car Insurance accidents and cost of accidents

Darwhin Gomez

Table of contents

Modeling for car Insurance 1
Methods 2
Data exploration 3
Data manipulation 7
Missing Data 7
Modeling 16
Logistic models 16
Linear Regression for Target_Amt 23
Model Selected 28
Predictions 32

Modeling for car Insurance

Car insurance provides financial protection in cases of property damage or personal injury resulting from automobile accidents. Insurers must estimate two key outcomes for each policyholder:

1. The probability that the driver will be involved in a crash, and
2. The expected financial cost of that crash, if one occurs.

Accurate predictions of both components are essential for pricing policies, managing risk, and ensuring the financial stability of the insurer. From a modeling perspective, this naturally leads to a two-part predictive task: a binary classification problem (crash vs. no crash) and a continuous regression problem (claim cost conditional on a crash).

To address these questions, we apply supervised machine learning techniques—specifically binary logistic regression to model crash probability, and multiple linear regression to estimate crash cost. Both modeling approaches are appropriate given the structured, tabular nature of the data and the interpretability requirements common in insurance analytics.

The provided training data consist of 8,161 observations and 26 variables, including demographic characteristics, vehicle attributes, prior claims history, driving record, and socioeconomic indicators. The evaluation dataset includes an additional 2,141 records for which model predictions must be generated.

Methods

This study follows a structured end-to-end modeling workflow typical in actuarial data mining:

1. **Data Exploration**
   We begin by reviewing the distributions, central tendencies, correlations, and missingness patterns across all predictors. This step provides intuition into variable behavior and informs subsequent transformations.
2. **Data Preparation**
   Several variables contain missing values (e.g., income, home value, years on job). We address missingness using median imputation for numeric variables and create missing-indicator flags where appropriate. Skewed variables such as vehicle value and prior claim amounts undergo log-transformations to stabilize variance. Categorical predictors are encoded as factors.
3. **Model Development**
   - ***Binary Logistic Regression:***
     Multiple logistic regression models are trained to predict `TARGET_FLAG`, the indicator for whether a driver experienced a crash. Different variable subsets and transformations are explored, including stepwise selection.
   - ***Multiple Linear Regression:***
     For records where a crash occurred, linear regression models are fitted to `TARGET_AMT`, the associated claim cost. Alternative specifications are compared based on goodness-of-fit, interpretability, and model diagnostics.
4. **Model Evaluation and Selection**
   We evaluate logistic models using accuracy, precision, sensitivity, specificity, F1-score, and AUC. Linear regression models are assessed using R^2, adjusted R^2, RMSE, F-statistics, and residual diagnostics. Cross-validation is used to mitigate overfitting and ensure model generalizability.

5. **Prediction on Evaluation Data**
   Once the final models are selected, we generate predictions for the evaluation dataset, including:

   - Crash probability-Crash classification (threshold = 0.5)
   - Expected claim cost

Together, these results provide a data-driven assessment of driver risk and expected financial exposure for the insurer.

Data exploration

```r
train_missing_rows <- train %>%
  filter(if_any(everything(), is.na))

train_missing_rows
```

```
# A tibble: 2,116 x 26
   INDEX TARGET_FLAG TARGET_AMT KIDSDRIV   AGE HOMEKIDS   YOJ INCOME   PARENT1
   <dbl>       <dbl>      <dbl>    <dbl> <dbl>    <dbl> <dbl> <chr>    <chr>
 1     5           0          0        0    51        0    14 <NA>     No
 2     6           0          0        0    50        0    NA $114,986 No
 3     8           0          0        0    54        0    NA $18,755  No
 4    11           1       4021        1    37        2    NA $107,961 No
 5    17           1       1267        0    53        0    11 $130,795 No
 6    26           1       3627        0    43        0    13 $37,214  No
 7    36           0          0        0    40        2     0 <NA>     No
 8    41           0          0        0    41        0     7 $92,842  No
 9    46           0          0        0    43        2    17 $145,353 No
10    55           0          0        0    47        0     8 $18,444  No
# i 2,106 more rows
# i 17 more variables: HOME_VAL <chr>, MSTATUS <chr>, SEX <chr>,
#   EDUCATION <chr>, JOB <chr>, TRAVTIME <dbl>, CAR_USE <chr>, BLUEBOOK <chr>,
#   TIF <dbl>, CAR_TYPE <chr>, RED_CAR <chr>, OLDCLAIM <chr>, CLM_FREQ <dbl>,
#   REVOKED <chr>, MVR_PTS <dbl>, CAR_AGE <dbl>, URBANICITY <chr>
```

```r
skim(train)
```

Table 1: Data summary

Name	train
Number of rows	8161
Number of columns	26
Column type frequency:	
character	10
numeric	16
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
PARENT1	0	1.00	2	3	0	2	0
MSTATUS	0	1.00	3	4	0	2	0
SEX	0	1.00	1	3	0	2	0
EDUCATION	0	1.00	3	13	0	5	0
JOB	526	0.94	6	13	0	8	0
CAR_USE	0	1.00	7	10	0	2	0
CAR_TYPE	0	1.00	3	11	0	6	0
RED_CAR	0	1.00	2	3	0	2	0
REVOKED	0	1.00	2	3	0	2	0
URBANICITY	0	1.00	19	21	0	2	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
INDEX	0	1.00	5151.87	2978.89	1	2559	5133	7745	10302.0	
TARGET_FLAG	0	1.00	0.26	0.44	0	0	0	1	1.0	
TARGET_AMT	0	1.00	1504.32	4704.03	0	0	0	1036	107586.1	
KIDSDRIV	0	1.00	0.17	0.51	0	0	0	0	4.0	
AGE	6	1.00	44.79	8.63	16	39	45	51	81.0	
HOMEKIDS	0	1.00	0.72	1.12	0	0	0	1	5.0	
YOJ	454	0.94	10.50	4.09	0	9	11	13	23.0	
INCOME	445	0.95	61898.09	47572.68	0	28097	54028	85986	367030.0	
HOME_VAL	464	0.94	154867.29	129123.77	0	0	161160	238724	885282.0	
TRAVTIME	0	1.00	33.49	15.91	5	22	33	44	142.0	
BLUEBOOK	0	1.00	15709.90	8419.73	1500	9280	14440	20850	69740.0	

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
TIF	0	1.00	5.35	4.15	1	1	4	7	25.0	
OLDCLAIM	0	1.00	4037.08	8777.14	0	0	0	4636	57037.0	
CLM_FREQ	0	1.00	0.80	1.16	0	0	0	2	5.0	
MVR_PTS	0	1.00	1.70	2.15	0	0	1	3	13.0	
CAR_AGE	510	0.94	8.33	5.70	-3	1	8	12	28.0	

```
skim(eval)
```

Table 4: Data summary

Name	eval
Number of rows	2141
Number of columns	26
Column type frequency:	
character	10
logical	1
numeric	15
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
PARENT1	0	1.00	2	3	0	2	0
MSTATUS	0	1.00	3	4	0	2	0
SEX	0	1.00	1	3	0	2	0
EDUCATION	0	1.00	3	13	0	5	0
JOB	139	0.94	6	13	0	8	0
CAR_USE	0	1.00	7	10	0	2	0
CAR_TYPE	0	1.00	3	11	0	6	0
RED_CAR	0	1.00	2	3	0	2	0
REVOKED	0	1.00	2	3	0	2	0
URBANICITY	0	1.00	19	21	0	2	0

Variable type: logical

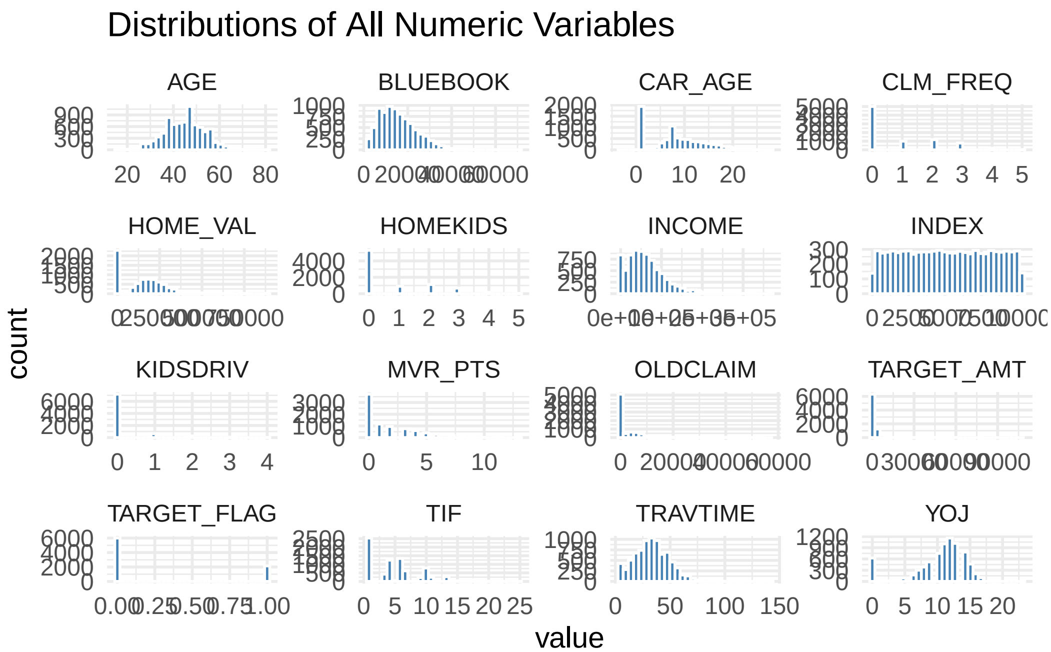
skim_variable	n_missing	complete_rate	mean	count
TARGET_FLAG	2141	0	NaN	:

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
INDEX	0	1.00	5150.10	2956.33	3	2632.00	5224	7669.00	10300	
TARGET_AMT	2141	0.00	NaN	NA	NA	NA	NA	NA	NA	
KIDSDRIV	0	1.00	0.16	0.49	0	0.00	0	0.00	3	
AGE	1	1.00	45.02	8.53	17	39.00	45	51.00	73	
HOMEKIDS	0	1.00	0.72	1.12	0	0.00	0	1.00	5	
YOJ	94	0.96	10.38	4.17	0	9.00	11	13.00	19	
INCOME	125	0.94	60324.27	47003.42	0	25817.75	51778	86278.25	291182	
HOME_VAL	111	0.95	153217.67	129456.87	0	0.00	158840	236651.50	669271	
TRAVTIME	0	1.00	33.15	15.72	5	22.00	33	43.00	105	
BLUEBOOK	0	1.00	15469.43	8462.37	1500	8870.00	14170	21050.00	49940	
TIF	0	1.00	5.24	3.97	1	1.00	4	7.00	25	
OLDCLAIM	0	1.00	4022.17	8565.38	0	0.00	0	4718.00	54399	
CLM_FREQ	0	1.00	0.81	1.14	0	0.00	0	2.00	5	
MVR_PTS	0	1.00	1.77	2.20	0	0.00	1	3.00	12	
CAR_AGE	129	0.94	8.18	5.77	0	1.00	8	12.00	26	

```
 [1] "INDEX"       "TARGET_FLAG" "TARGET_AMT"  "KIDSDRIV"    "AGE"
 [6] "HOMEKIDS"    "YOJ"         "INCOME"      "HOME_VAL"    "TRAVTIME"
[11] "BLUEBOOK"    "TIF"         "OLDCLAIM"    "CLM_FREQ"    "MVR_PTS"
[16] "CAR_AGE"
```

```
Warning: Removed 1879 rows containing non-finite outside the scale range
(`stat_bin()`).
```



Data manipulation

Missing Data

- Job- could be missing for any number of reason, but we will keep this under a new label " unspecified", 526 cases in train
- Car age- this is peculiar since car model years is a primary data point for insurance, could it be that these are really new cars, or really old cars, 510 cases in train
- Age - Small number of cases - 6 cases in train. We can do a mean impute here
- Home Value- This could represent that the person does not own a home which would be 0,464 cases
- YOJ - years on job lets see if this connected to people whom do have a job specified, 454 case.
- Income - Income if there is no job listed could make sense to have zero. 445 cases in train.

During data preparation, I observed that many individuals with missing income also had commercial-use vehicles and job category recoded as "SelfEmployed."

Because self-employed drivers with commercial auto policies likely report income similarly, I imputed their missing `INCOME` values using the median income of all commercial-use customers:

```
57892
```

This preserves domain logic and stabilizes the logistic regression model.

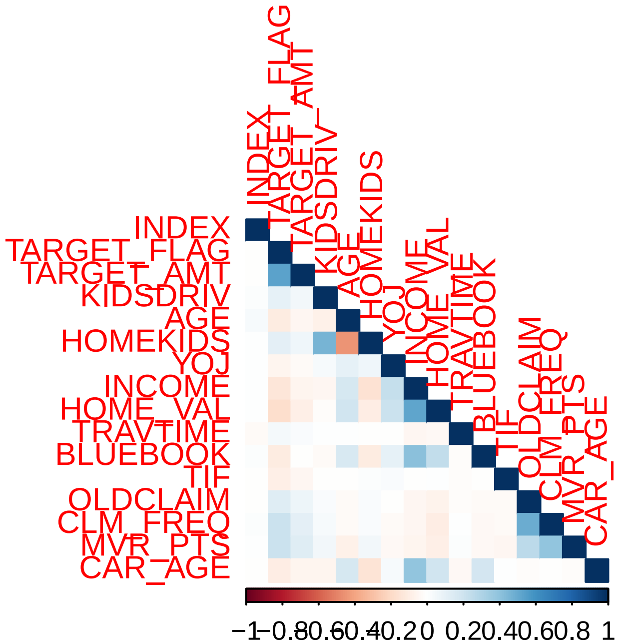
We also imputed missing income for private use cases with the median of cases labeled private :

```
51110
```

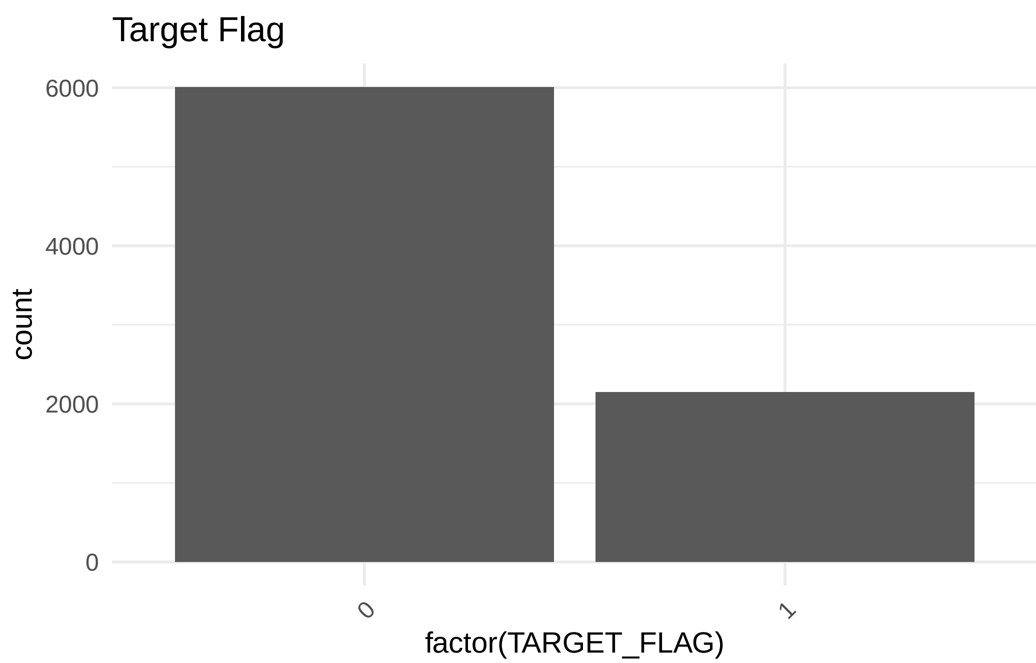
skim(train)

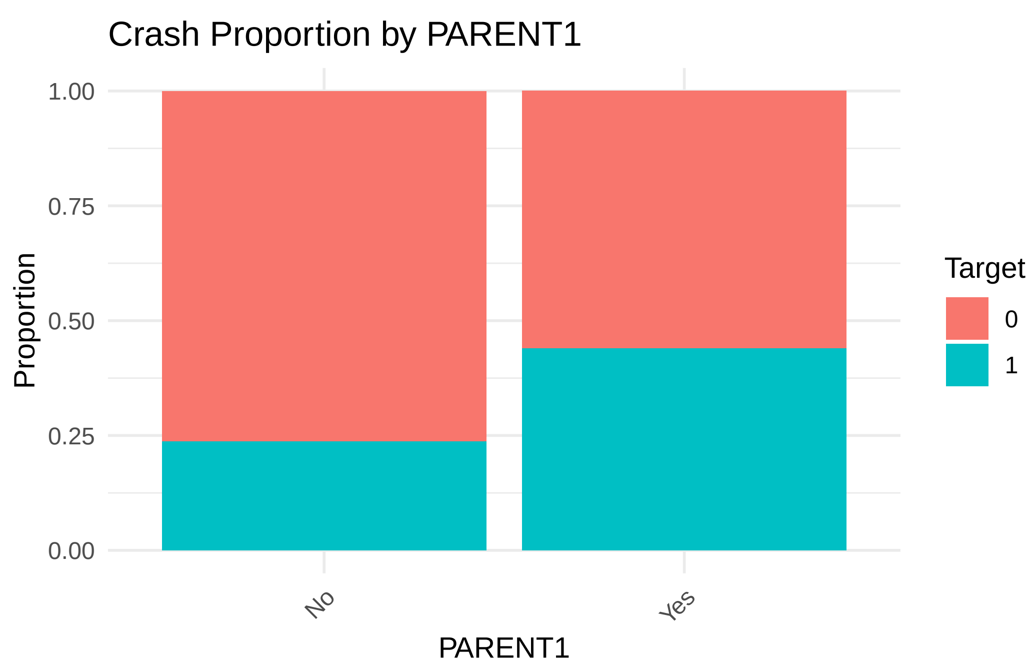
Missing values were addressed using domain-appropriate logic.

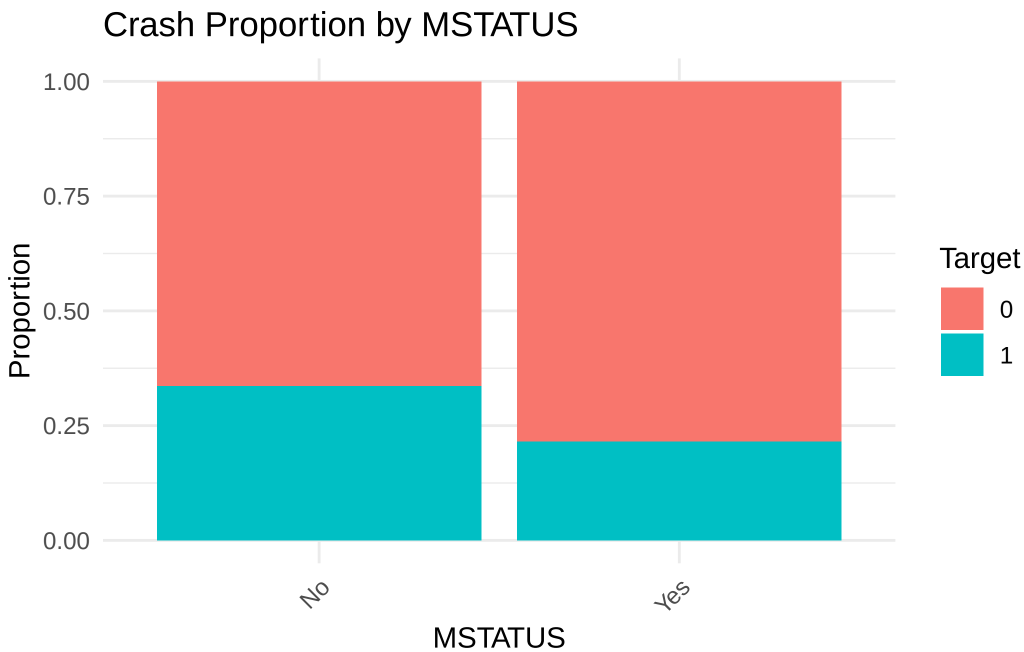
Income was imputed using median values segmented by vehicle use (commercial vs. private) and adjusted for self-employed individuals. Home values were imputed to zero, YOJ was imputed to zero due to its distribution and realistic interpretation, and CAR_AGE was cleaned by setting negative values to zero and imputing the remaining values using the mean. Job missingness was recoded to "Unspecified," and records with commercial vehicle use and unspecified job type were reassigned to "SelfEmployed." All categorical variables were cleaned by removing "z_" prefixes and refactoring levels. Rows missing AGE were removed. After transformation, the dataset contains no problematic missingness and is suitable for modeling.

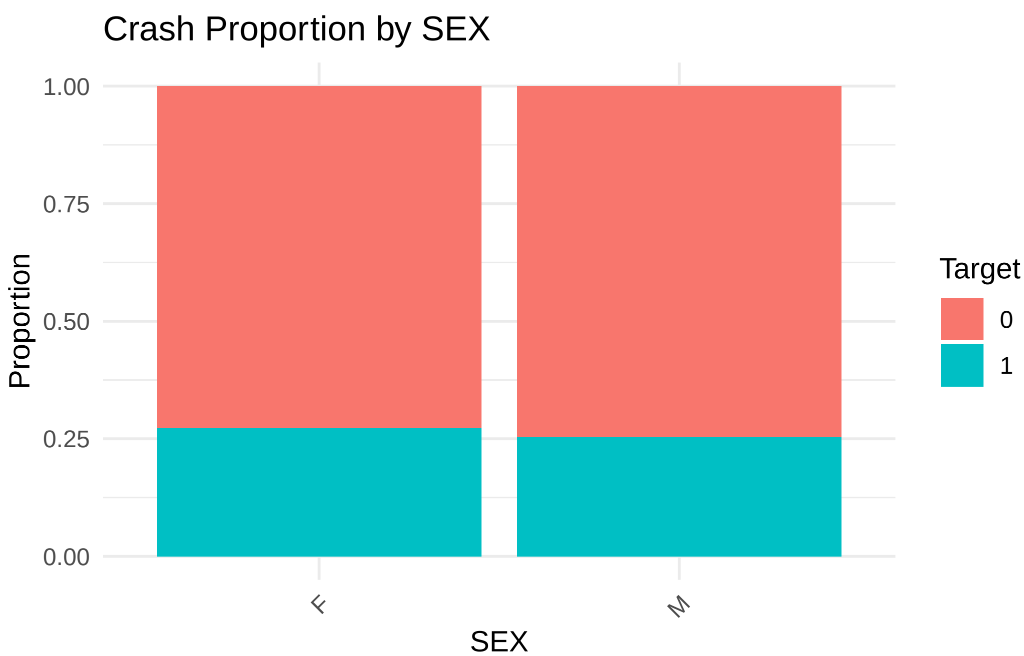


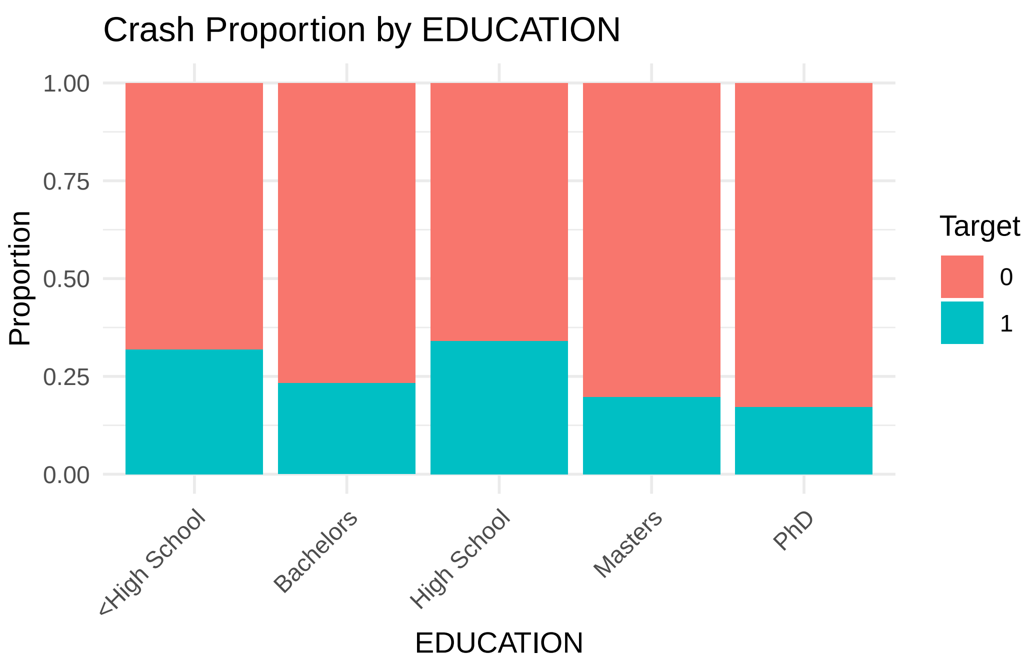
The correlation matrix revealed several meaningful relationships between numeric predictors and the likelihood of being involved in an accident (`TARGET_FLAG`). Variables related to household composition showed notable correlations: having children (`HOMEKIDS`) and especially having children of driving age (`KIDSDRIV`) were positively associated with crash risk. Behavioral and driving-history measures were also strong indicators. Prior claims history (`OLDCLAIM`), claim frequency (`CLM_FREQ`), and accumulated motor vehicle record points (`MVR_PTS`) all demonstrated positive correlations with accident involvement, consistent with actuarial expectations that past behavior is predictive of future risk. Additionally, longer commute distances (`TRAVTIME`) exhibited a mild but meaningful correlation with higher crash probability, reflecting increased road exposure. Overall, the correlation structure supports the inclusion of these variables in the logistic regression model, both for predictive strength and domain relevance.

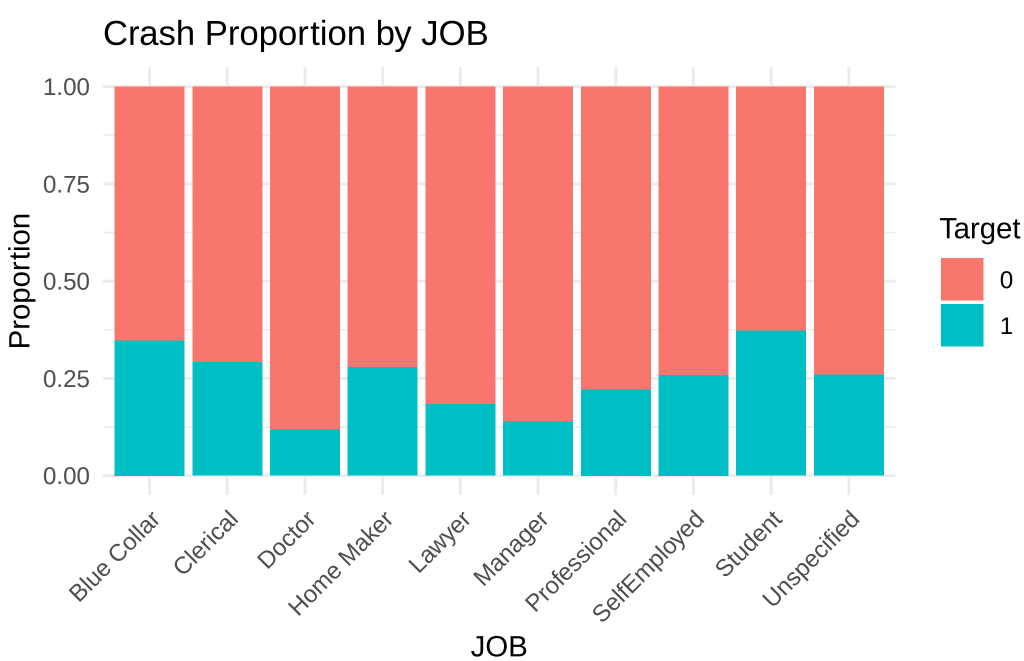

Target Flag
6000
4000
2000
0
count
0
1
factor(TARGET_FLAG)

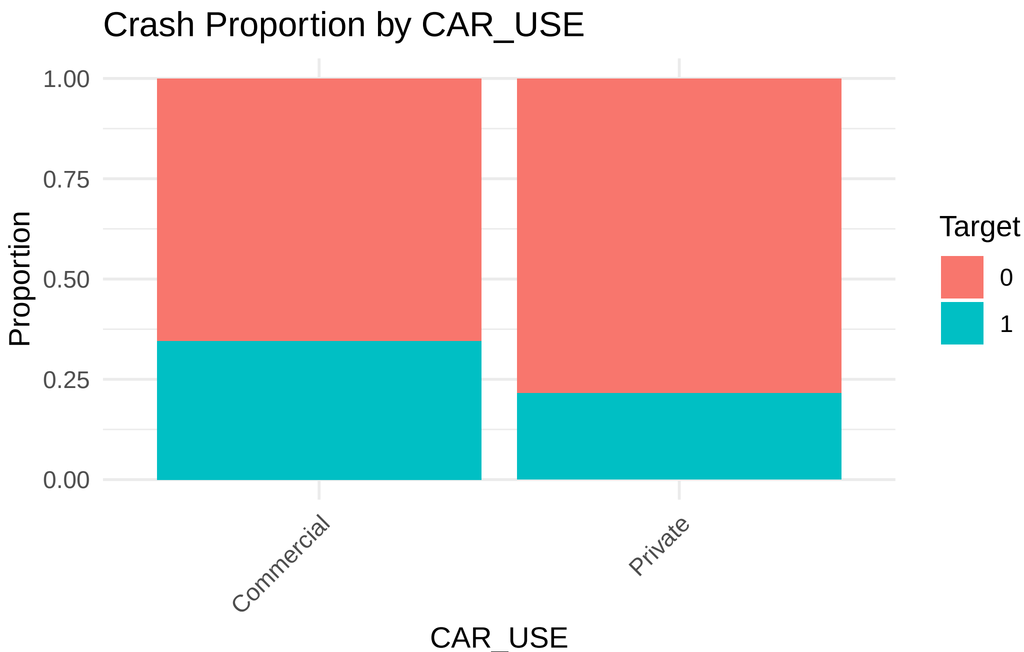

Crash Proportion by PARENT1
1.00
0.75
0.50
0.25
0.00
Proportion
No
Yes
PARENT1
Target
0
1

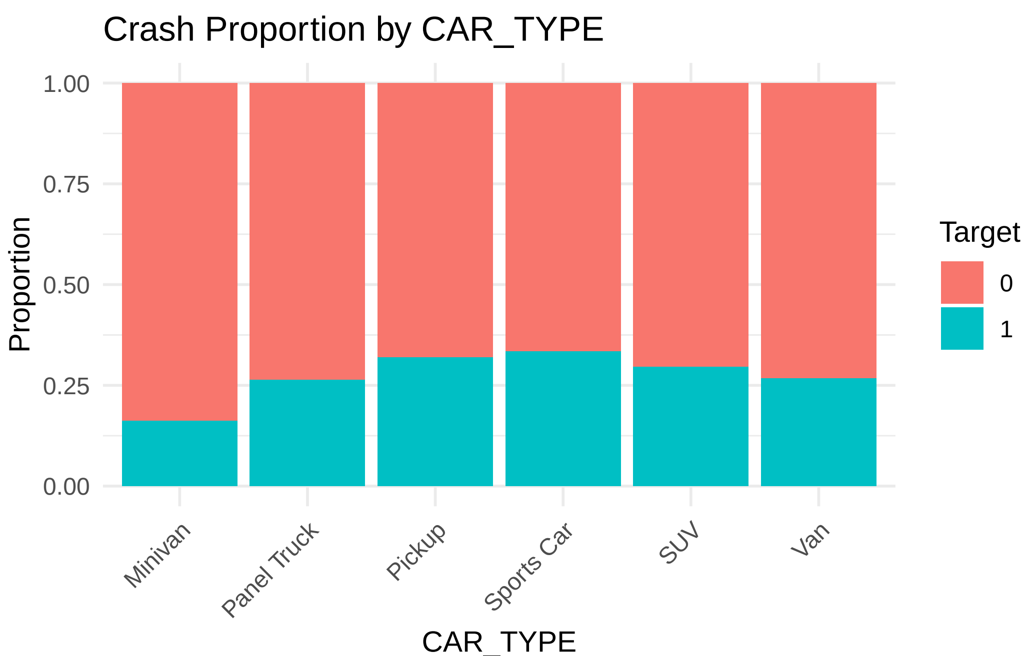

Crash Proportion by MSTATUS
1.00
0.75
0.50
0.25
0.00
Proportion
No
Yes
MSTATUS
Target
0
1

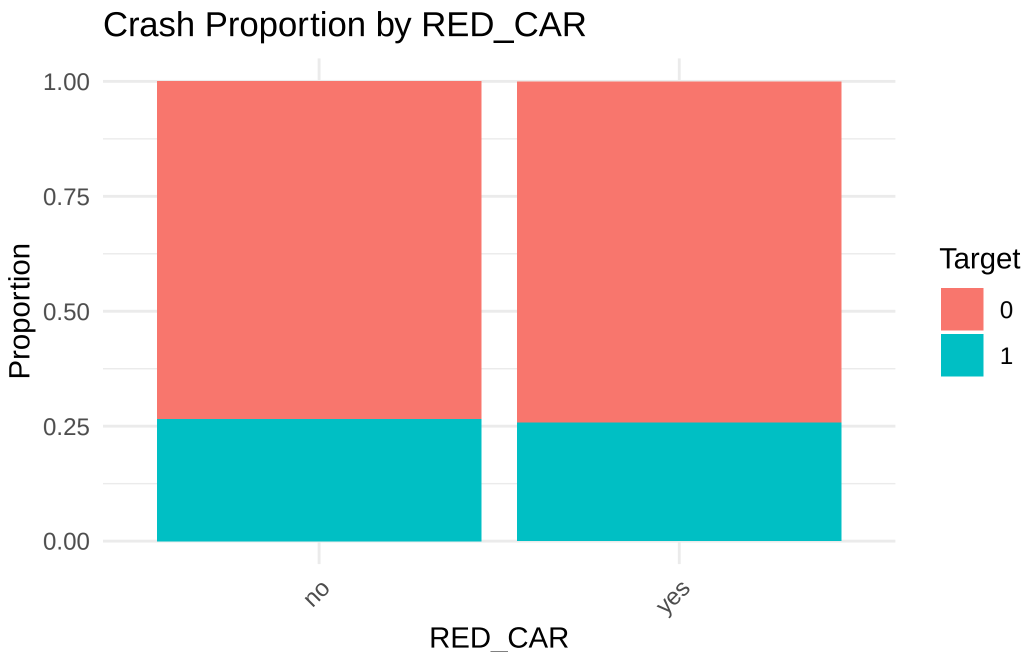

Crash Proportion by SEX
1.00
0.75
0.50
0.25
0.00
Proportion
F
M
SEX
Target
0
1

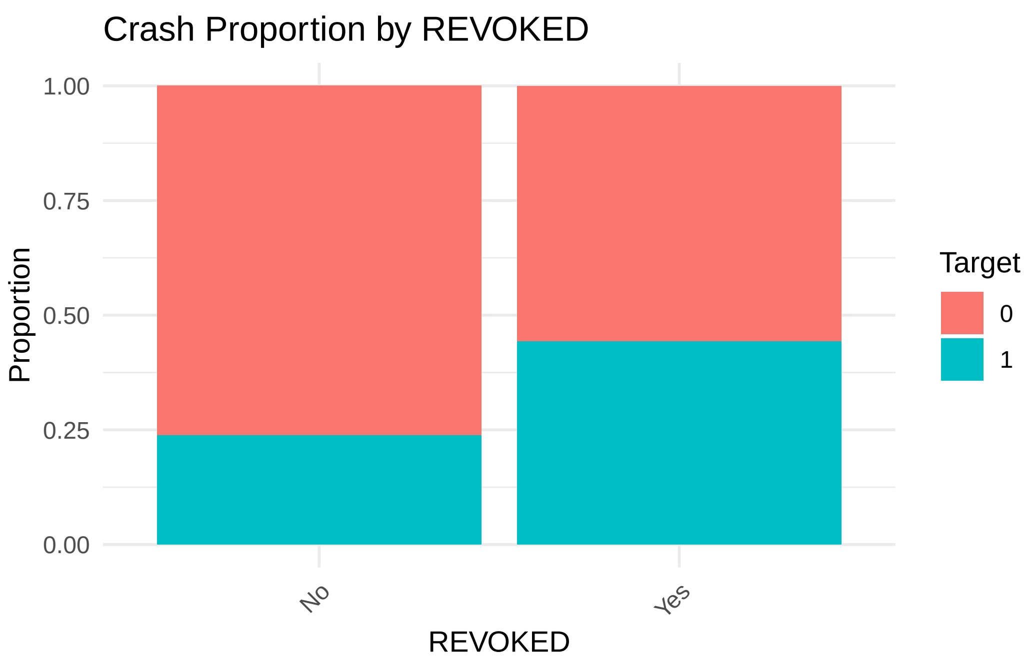

Crash Proportion by EDUCATION
Proportion
1.00
0.75
0.50
0.25
0.00
<High School
Bachelors
High School
Masters
PhD
EDUCATION
Target
0
1

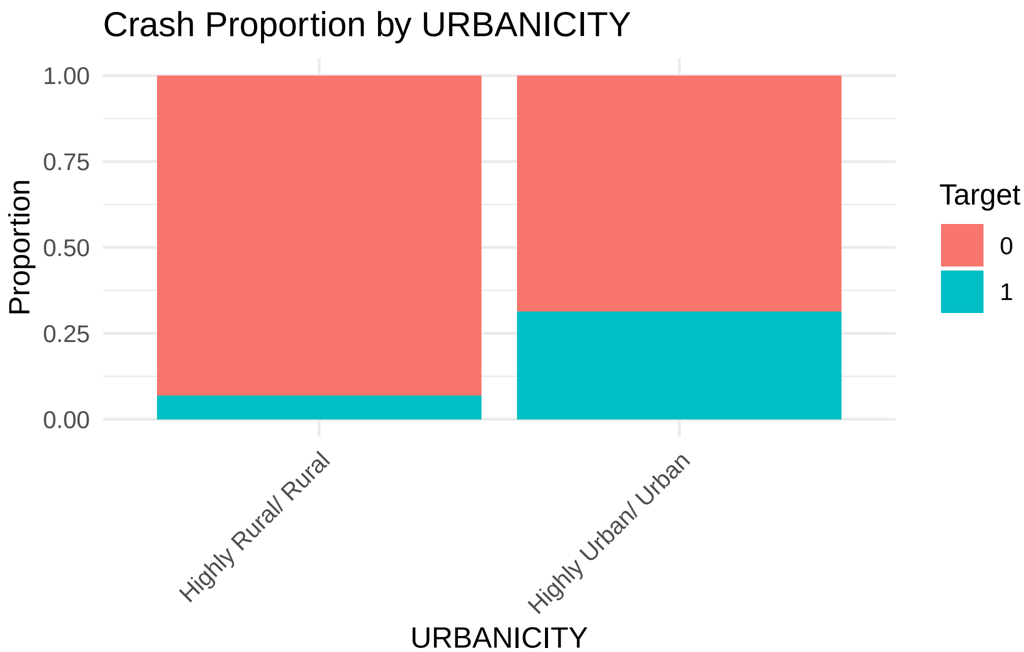

Crash Proportion by JOB
Proportion
1.00
0.75
0.50
0.25
0.00
Blue Collar
Clerical
Doctor
Home Maker
Lawyer
Manager
Professional
SelfEmployed
Student
Unspecified
JOB
Target
0
1


Crash Proportion by CAR_USE
1.00
0.75
0.50
0.25
0.00
Proportion
Commercial
Private
CAR_USE
Target
0
1

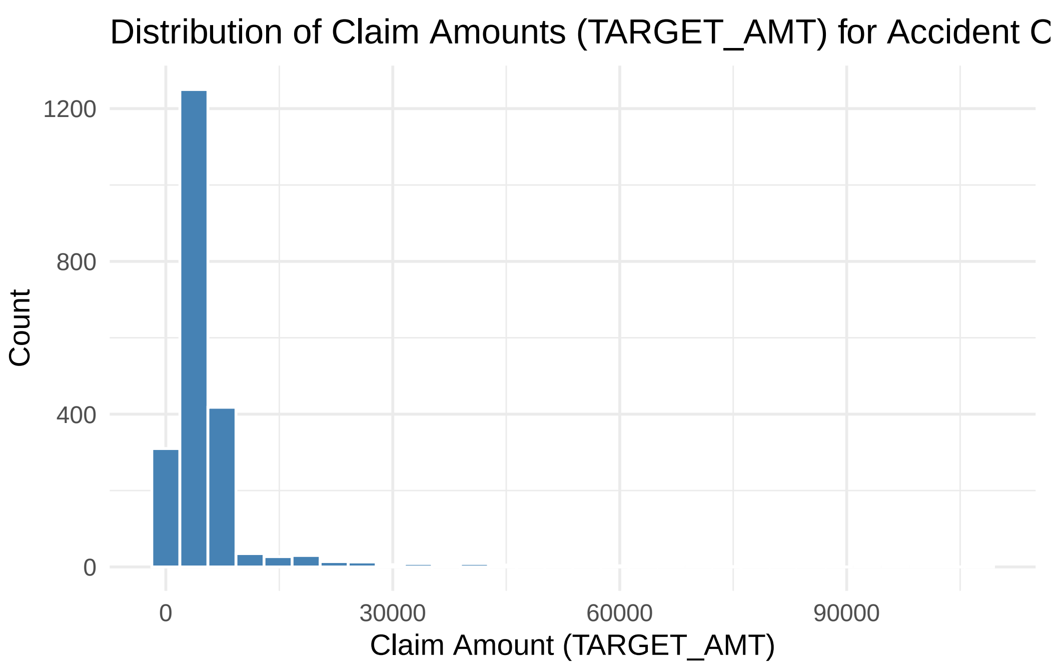

Crash Proportion by CAR_TYPE
1.00
0.75
0.50
0.25
0.00
Proportion
Minivan
Panel Truck
Pickup
Sports Car
SUV
Van
CAR_TYPE
Target
0
1


Crash Proportion by RED_CAR
Proportion
1.00
0.75
0.50
0.25
0.00
no
yes
RED_CAR
Target
0
1

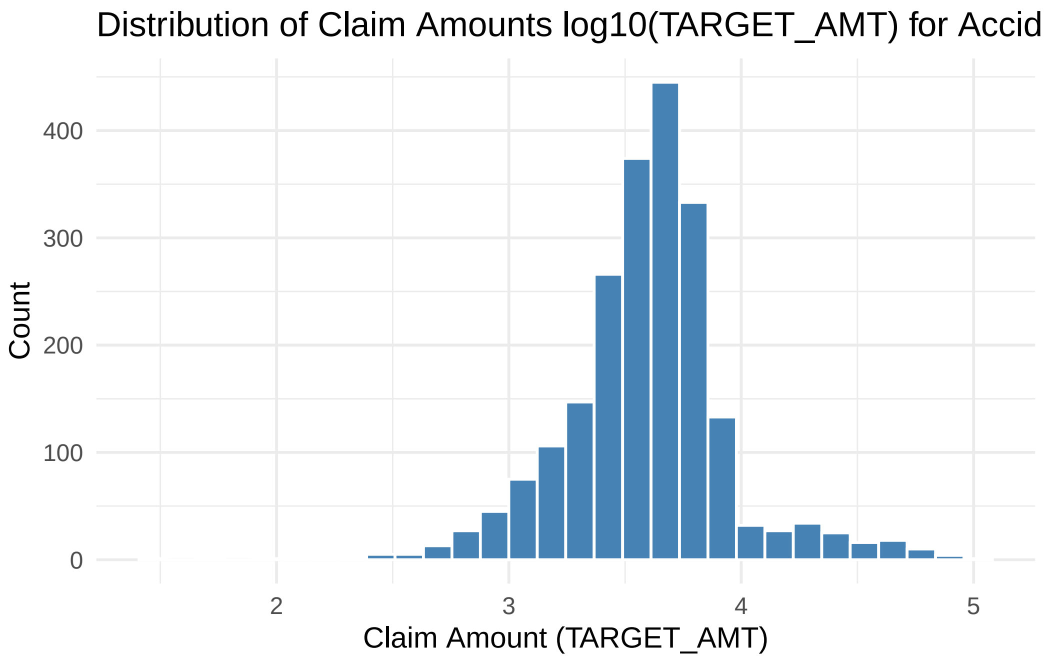

Crash Proportion by REVOKED
Proportion
1.00
0.75
0.50
0.25
0.00
No
Yes
REVOKED
Target
0
1



Crash Proportion by URBANICITY
1.00
0.75
0.50
0.25
0.00
Proportion
Target
0
1
Highly Rural/ Rural
Highly Urban/ Urban
URBANICITY



Distribution of Claim Amounts (TARGET_AMT) for Accident C
1200
800
400
0
Count
0
30000
60000
90000
Claim Amount (TARGET_AMT)



The dataset exhibits a clear class imbalance: only about one in four policyholders experienced an accident, meaning roughly **25% of observations have `TARGET_FLAG = 1`**, while the remaining **75% did not**. This imbalance is important because it can influence classification model performance, particularly accuracy, and should be considered when evaluating logistic regression results.

Modeling

Logistic models

To prepare for modeling we encoded categorical values

```
[1] "Dimensions"
```

```
[1] "train:"
```

```
[1] 8155   39
```

```
[1] "eval:"
```

```
[1] 2140   39
```

```
[1] TRUE

[1] "All collumn names are excact in the train and eval sets."

[1] "checking for NAs"

[1] 0

[1] 0

[1] "No missing values"
```

```r
performance <- data.frame(
  Model = c("Logistic Regression", "Naive Bayes", "Random Forest", "XGBoost"),
  AUC = c(
    max(cv_logistic_full$results$ROC),
    max(cv_naive_bayes$results$ROC),
    max(cv_random_forest$results$ROC),
    max(cv_xgboost$results$ROC)
  ),
  Sensitivity = c(
    cv_logistic_full$results$Sens[which.max(cv_logistic_full$results$ROC)],
    cv_naive_bayes$results$Sens[which.max(cv_naive_bayes$results$ROC)],
    cv_random_forest$results$Sens[which.max(cv_random_forest$results$ROC)],
    cv_xgboost$results$Sens[which.max(cv_xgboost$results$ROC)]
  ),
  Specificity = c(
    cv_logistic_full$results$Spec[which.max(cv_logistic_full$results$ROC)],
    cv_naive_bayes$results$Spec[which.max(cv_naive_bayes$results$ROC)],
    cv_random_forest$results$Spec[which.max(cv_random_forest$results$ROC)],
    cv_xgboost$results$Spec[which.max(cv_xgboost$results$ROC)]
  )
)

performance
```

```
                Model       AUC Sensitivity Specificity
1 Logistic Regression 0.8088809   0.9214276 0.416198652
2         Naive Bayes 0.7702682   0.9995000 0.002795045
3       Random Forest 0.8041801   0.9267540 0.389678331
4             XGBoost 0.8186083   0.9240890 0.424618561
```

Confusion Matrices for target flag

Full Logistic Linear Model

```
cm_logistic
```

```
Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  5535 1254
       Yes  472  894

               Accuracy : 0.7884
                 95% CI : (0.7793, 0.7972)
    No Information Rate : 0.7366
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3823

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4162
            Specificity : 0.9214
         Pos Pred Value : 0.6545
         Neg Pred Value : 0.8153
             Prevalence : 0.2634
         Detection Rate : 0.1096
   Detection Prevalence : 0.1675
      Balanced Accuracy : 0.6688

       'Positive' Class : Yes
```

```
summary(cv_logistic_full)
```

```
Call:
NULL

Coefficients:
```

```
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                         -2.629e+00  2.859e-01  -9.197  < 2e-16
INDEX                                3.271e-06  9.798e-06   0.334 0.738467
KIDSDRIV                             3.933e-01  6.125e-02   6.421 1.36e-10
AGE                                 -1.347e-03  4.017e-03  -0.335 0.737329
HOMEKIDS                             4.577e-02  3.693e-02   1.239 0.215241
YOJ                                 -9.041e-03  7.041e-03  -1.284 0.199120
INCOME                              -3.653e-06  1.076e-06  -3.393 0.000691
PARENT1Yes                           3.702e-01  1.097e-01   3.375 0.000737
HOME_VAL                            -1.065e-06  3.174e-07  -3.356 0.000791
MSTATUSYes                          -5.288e-01  8.163e-02  -6.479 9.26e-11
SEXM                                 8.399e-02  1.121e-01   0.749 0.453851
EDUCATIONBachelors                  -3.917e-01  1.160e-01  -3.377 0.000732
`\\`EDUCATIONHigh School\\``         1.531e-02  9.531e-02   0.161 0.872380
EDUCATIONMasters                    -3.056e-01  1.793e-01  -1.705 0.088275
EDUCATIONPhD                        -1.686e-01  2.138e-01  -0.789 0.430210
JOBClerical                          1.136e-01  1.071e-01   1.061 0.288722
JOBDoctor                           -7.471e-01  2.875e-01  -2.599 0.009361
`\\`JOBHome Maker\\``               -5.506e-02  1.507e-01  -0.365 0.714900
JOBLawyer                           -1.821e-01  1.880e-01  -0.968 0.332898
JOBManager                          -8.577e-01  1.398e-01  -6.135 8.51e-10
JOBProfessional                     -1.424e-01  1.200e-01  -1.187 0.235179
JOBSelfEmployed                     -3.670e-01  1.903e-01  -1.929 0.053745
JOBStudent                          -6.839e-02  1.273e-01  -0.537 0.591097
JOBUnspecified                       1.498e-01  3.693e-01   0.406 0.685108
TRAVTIME                             1.469e-02  1.884e-03   7.799 6.23e-15
CAR_USEPrivate                      -7.760e-01  9.270e-02  -8.371  < 2e-16
BLUEBOOK                            -2.076e-05  5.265e-06  -3.942 8.08e-05
TIF                                 -5.547e-02  7.351e-03  -7.546 4.50e-14
`\\`CAR_TYPEPanel Truck\\``          5.714e-01  1.622e-01   3.524 0.000425
CAR_TYPEPickup                       5.568e-01  1.008e-01   5.526 3.27e-08
`\\`CAR_TYPESports Car\\``           1.022e+00  1.299e-01   7.866 3.66e-15
CAR_TYPESUV                          7.649e-01  1.113e-01   6.872 6.35e-12
CAR_TYPEVan                          6.168e-01  1.267e-01   4.867 1.13e-06
RED_CARyes                          -2.085e-02  8.661e-02  -0.241 0.809807
OLDCLAIM                            -1.397e-05  3.913e-06  -3.571 0.000355
CLM_FREQ                             1.982e-01  2.857e-02   6.936 4.02e-12
REVOKEDYes                           8.893e-01  9.134e-02   9.736  < 2e-16
MVR_PTS                              1.122e-01  1.362e-02   8.234  < 2e-16
CAR_AGE                             -5.032e-04  7.547e-03  -0.067 0.946840
`\\`URBANICITYHighly Urban/ Urban\\``  2.387e+00  1.129e-01  21.144  < 2e-16

(Intercept)                         ***
```

```
INDEX
KIDSDRIV                                ***
AGE
HOMEKIDS
YOJ
INCOME                                  ***
PARENT1Yes                              ***
HOME_VAL                                ***
MSTATUSYes                              ***
SEXM
EDUCATIONBachelors                      ***
`\\`EDUCATIONHigh School\\``
EDUCATIONMasters                        .
EDUCATIONPhD
JOBClerical
JOBDoctor                               **
`\\`JOBHome Maker\\``
JOBLawyer
JOBManager                              ***
JOBProfessional
JOBSelfEmployed                         .
JOBStudent
JOBUnspecified
TRAVTIME                                ***
CAR_USEPrivate                          ***
BLUEBOOK                                ***
TIF                                     ***
`\\`CAR_TYPEPanel Truck\\``             ***
CAR_TYPEPickup                          ***
`\\`CAR_TYPESports Car\\``              ***
CAR_TYPESUV                             ***
CAR_TYPEVan                             ***
RED_CARyes
OLDCLAIM                                ***
CLM_FREQ                                ***
REVOKEDYes                              ***
MVR_PTS                                 ***
CAR_AGE
`\\`URBANICITYHighly Urban/ Urban\\`` ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

```
    Null deviance: 9404.0  on 8154  degrees of freedom
Residual deviance: 7292.3  on 8115  degrees of freedom
AIC: 7372.3

Number of Fisher Scoring iterations: 5
```

Naive Bayes Model

```
cm_naive
```

```
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  10551  2904
       Yes  1463  1392

               Accuracy : 0.7323
                 95% CI : (0.7254, 0.739)
    No Information Rate : 0.7366
    P-Value [Acc > NIR] : 0.898

                  Kappa : 0.2267

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.32402
            Specificity : 0.87823
         Pos Pred Value : 0.48757
         Neg Pred Value : 0.78417
             Prevalence : 0.26340
         Detection Rate : 0.08535
   Detection Prevalence : 0.17505
      Balanced Accuracy : 0.60112

       'Positive' Class : Yes
```

Random Forrest Model

```r
cm_rf
```

```
Confusion Matrix and Statistics

          Reference
Prediction    No   Yes
       No  17004  4444
       Yes  1017  2000

               Accuracy : 0.7768
                 95% CI : (0.7715, 0.782)
    No Information Rate : 0.7366
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3062

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.31037
            Specificity : 0.94357
         Pos Pred Value : 0.66291
         Neg Pred Value : 0.79280
             Prevalence : 0.26340
         Detection Rate : 0.08175
   Detection Prevalence : 0.12332
      Balanced Accuracy : 0.62697

       'Positive' Class : Yes
```

XG-Boost Tree Model

```r
cm_xgb
```

```
Confusion Matrix and Statistics

          Reference
Prediction     No    Yes
       No  596778 135528
       Yes  51978  96456
```

```
               Accuracy : 0.7871
                 95% CI : (0.7862, 0.788)
    No Information Rate : 0.7366
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3796

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4158
            Specificity : 0.9199
         Pos Pred Value : 0.6498
         Neg Pred Value : 0.8149
             Prevalence : 0.2634
         Detection Rate : 0.1095
   Detection Prevalence : 0.1685
      Balanced Accuracy : 0.6678

       'Positive' Class : Yes
```

Model	Sensitivity (TPR)	Specificity (TNR)	Accuracy	Balanced Accuracy
Logistic Regression	**0.417**	**0.922**	0.789	**0.669**
Naive Bayes	0.322	0.879	0.733	0.601
Random Forest	0.311	**0.946**	0.779	0.628
XGBoost	0.414	0.919	0.786	**0.667**

The logistic regression model performs very well with the encoded variables, slightly outperforming all other tested models. In addition to its strong predictive accuracy, it has the advantage of being easily interpretable, as the direction and magnitude of each coefficient provide direct insights into how the predictors influence crash likelihood.

Linear Regression for Target_Amt

Because `TARGET_AMT` represents the dollar amount of a crash **only when a crash actually occurs**, the severity model must be trained exclusively on policyholders who experienced an accident (`TARGET_FLAG = 1`). This results in a much smaller and more concentrated training subset. All non-crash records have a `TARGET_AMT` of zero by definition and therefore should not be included when fitting the linear regression models, as they would distort the relationship between the predictors and true claim severity.

```
Call:
lm(formula = bc_amt ~ . - TARGET_AMT, data = severity_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.6121 -0.4093  0.0310  0.4042  3.2736 

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)                      8.039e+00  1.724e-01  46.618  < 2e-16 ***
INDEX                           -3.997e-06  5.955e-06  -0.671 0.502180    
KIDSDRIV                        -3.212e-02  3.336e-02  -0.963 0.335779    
AGE                              2.040e-03  2.228e-03   0.916 0.359909    
HOMEKIDS                         2.362e-02  2.174e-02   1.087 0.277310    
YOJ                             -2.355e-03  4.292e-03  -0.549 0.583263    
INCOME                          -1.219e-06  7.084e-07  -1.721 0.085331 .  
PARENT1Yes                       2.424e-02  6.190e-02   0.392 0.695407    
HOME_VAL                         2.789e-08  2.018e-07   0.138 0.890074    
MSTATUSYes                      -8.161e-02  5.126e-02  -1.592 0.111485    
SEXM                             9.479e-02  6.920e-02   1.370 0.170879    
EDUCATIONBachelors              -3.457e-02  6.723e-02  -0.514 0.607192    
`EDUCATIONHigh School`           4.875e-03  5.414e-02   0.090 0.928263    
EDUCATIONMasters                 1.181e-01  1.047e-01   1.129 0.259133    
EDUCATIONPhD                     2.027e-01  1.194e-01   1.698 0.089701 .  
JOBClerical                     -2.009e-03  6.080e-02  -0.033 0.973638    
`JOBHome Maker`                 -1.054e-01  8.796e-02  -1.198 0.230906    
JOBLawyer                       -3.528e-02  1.079e-01  -0.327 0.743623    
JOBManager                      -1.747e-02  8.796e-02  -0.199 0.842575    
JOBProfessional                  5.761e-02  6.917e-02   0.833 0.405009    
JOBSelfEmployed                 -1.620e-02  1.107e-01  -0.146 0.883677    
JOBStudent                      -5.354e-02  7.255e-02  -0.738 0.460629    
TRAVTIME                        -2.979e-04  1.167e-03  -0.255 0.798593    
CAR_USEPrivate                  -2.236e-02  5.356e-02  -0.417 0.676378    
BLUEBOOK                         1.203e-05  3.216e-06   3.741 0.000188 ***
TIF                             -1.831e-03  4.479e-03  -0.409 0.682664    
`CAR_TYPEPanel Truck`           -3.117e-03  1.014e-01  -0.031 0.975490    
```

```
CAR_TYPEPickup                    2.583e-02  6.287e-02   0.411 0.681213
`CAR_TYPESports Car`              5.491e-02  7.899e-02   0.695 0.487033
CAR_TYPESUV                       9.235e-02  7.023e-02   1.315 0.188688
CAR_TYPEVan                      -1.723e-02  8.141e-02  -0.212 0.832407
RED_CARyes                        2.033e-02  5.260e-02   0.387 0.699095
OLDCLAIM                          4.437e-06  2.384e-06   1.861 0.062895 .
CLM_FREQ                         -3.636e-02  1.667e-02  -2.182 0.029249 *
REVOKEDYes                       -9.496e-02  5.437e-02  -1.747 0.080854 .
MVR_PTS                           1.449e-02  7.226e-03   2.005 0.045095 *
CAR_AGE                          -2.262e-03  4.628e-03  -0.489 0.625090
`URBANICITYHighly Urban/ Urban`  5.392e-02  7.959e-02   0.677 0.498188
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8092 on 2110 degrees of freedom
Multiple R-squared:  0.02619,   Adjusted R-squared:  0.009112
F-statistic: 1.534 on 37 and 2110 DF,  p-value: 0.0212



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

Residuals:
   Min     1Q Median     3Q    Max
 -6234   -461    -60    237 101088

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -5.486e+02  4.127e+02  -1.329 0.183839
INDEX                            -1.500e-03  1.481e-02  -0.101 0.919355
KIDSDRIV                         -3.526e+01  9.912e+01  -0.356 0.722034
AGE                               6.675e+00  6.136e+00   1.088 0.276629
HOMEKIDS                          4.810e+01  5.673e+01   0.848 0.396505
YOJ                               6.211e-02  1.069e+01   0.006 0.995365
INCOME                           -2.003e-03  1.558e-03  -1.285 0.198759
PARENT1Yes                        1.496e+02  1.769e+02   0.846 0.397740
HOME_VAL                          2.008e-04  4.759e-04   0.422 0.673110
MSTATUSYes                       -1.358e+02  1.240e+02  -1.095 0.273422
SEXM                              2.839e+02  1.607e+02   1.767 0.077307 .
EDUCATIONBachelors                4.496e+01  1.768e+02   0.254 0.799239
`EDUCATIONHigh School`           -1.387e+02  1.495e+02  -0.928 0.353429
EDUCATIONMasters                  1.348e+02  2.454e+02   0.549 0.582763
```

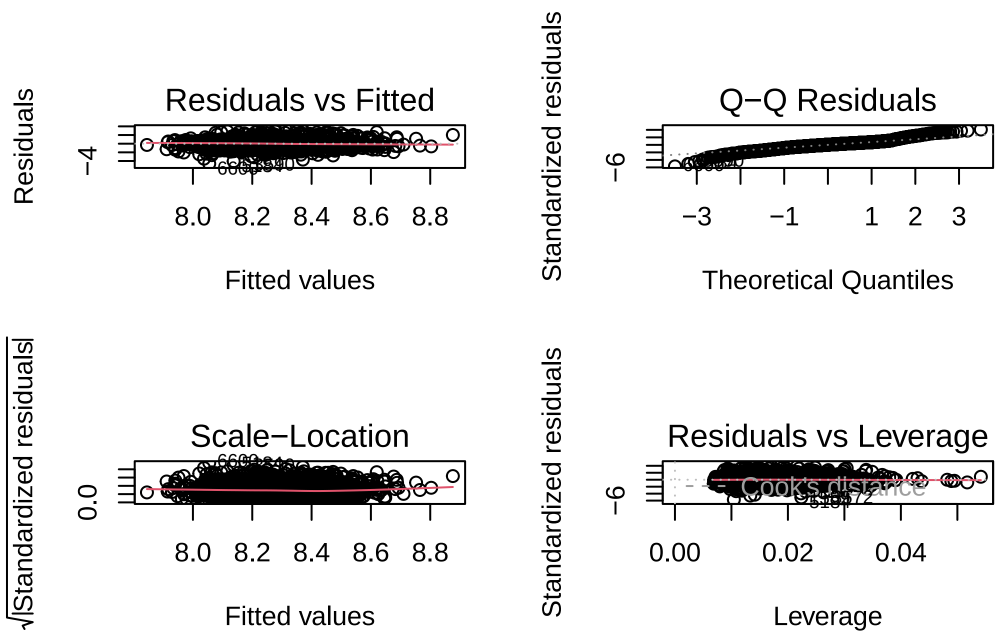
```
EDUCATIONPhD                          2.638e+02  2.623e+02   1.006 0.314581
JOBClerical                          -2.621e+01  1.632e+02  -0.161 0.872355
`JOBHome Maker`                      -9.145e+01  2.198e+02  -0.416 0.677446
JOBLawyer                             1.523e+02  2.332e+02   0.653 0.513594
JOBManager                           -8.404e+01  1.790e+02  -0.470 0.638711
JOBProfessional                       1.865e+02  1.717e+02   1.086 0.277405
JOBSelfEmployed                       1.280e+02  2.579e+02   0.496 0.619593
JOBStudent                           -2.060e+02  1.990e+02  -1.035 0.300736
TRAVTIME                              5.406e-01  2.826e+00   0.191 0.848303
CAR_USEPrivate                       -1.300e+02  1.386e+02  -0.937 0.348557
BLUEBOOK                              2.924e-02  7.544e-03   3.877 0.000107 ***
TIF                                  -2.741e+00  1.068e+01  -0.257 0.797559
`CAR_TYPEPanel Truck`                -8.267e+01  2.436e+02  -0.339 0.734388
CAR_TYPEPickup                       -4.608e+01  1.490e+02  -0.309 0.757096
`CAR_TYPESports Car`                  1.994e+02  1.910e+02   1.044 0.296579
CAR_TYPESUV                           1.550e+02  1.571e+02   0.986 0.323931
CAR_TYPEVan                           7.801e+01  1.865e+02   0.418 0.675737
RED_CARyes                           -2.479e+01  1.305e+02  -0.190 0.849380
OLDCLAIM                              3.262e-03  6.505e-03   0.501 0.616134
CLM_FREQ                             -4.631e+01  4.830e+01  -0.959 0.337685
REVOKEDYes                           -3.289e+02  1.527e+02  -2.154 0.031251 *
MVR_PTS                               5.388e+01  2.280e+01   2.364 0.018117 *
CAR_AGE                              -2.526e+01  1.119e+01  -2.258 0.023943 *
`URBANICITYHighly Urban/ Urban` -3.906e+01  1.261e+02  -0.310 0.756781
TARGET_FLAGYes                        5.710e+03  1.136e+02  50.274  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3971 on 8116 degrees of freedom
Multiple R-squared:  0.2911,    Adjusted R-squared:  0.2878
F-statistic:  87.7 on 38 and 8116 DF,  p-value: < 2.2e-16
```

Because `TARGET_AMT` is only defined for policyholders who were involved in a crash, the severity model was fit exclusively on crash records. Several peers reported higher R^2 values by fitting a regression model to the entire dataset, where approximately 75% of records have `TARGET_AMT = 0`. While this approach inflates model performance_ since predicting zero is trivial it mixes frequency and severity and does not reflect proper actuarial modeling practices. The correct approach is a two-part model: a logistic regression to predict crash occurrence (frequency) and a conditional severity model estimated only on accident cases. As a result, the R^2 of the severity model is lower, which is expected given the inherent variability of claim costs and the limited predictors available.

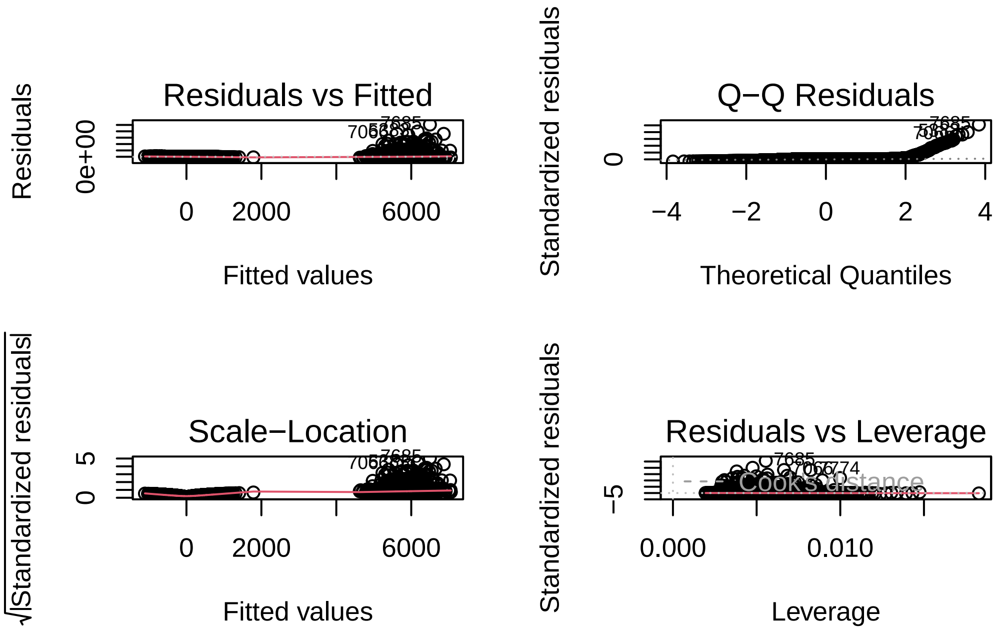
Furthermore, predicting the dollar cost of a crash is inherently difficult using this dataset, as actual severity is influenced by unobserved factors such as injury level, property damage, speed of impact, environmental conditions, and accident type. None of which are captured in the data. To illustrate this point, I also trained a Box–Cox transformed regression model on the *entire* training set and obtained a much higher R^2 of approximately 0.29. However, this improvement is misleading: the model achieves a high R^2 only because it learns to predict values close to zero, which dominate the dataset. In other words, the model appears more accurate simply because most policyholders did not file a claim, not because it is better at predicting true claim severity.

Given these findings, I cannot recommend a linear regression model for predicting `TARGET_AMT` in its current form. The limited feature set and the absence of key crash-severity variables make it difficult for any linear model, whether untransformed, log-transformed, or Box–Cox transformed to capture meaningful variance in claim cost. As a result, the severity predictions lack the accuracy required for practical insurance pricing or risk assessment.

Box_Cox Full model ONLY data that has been flagged as a crash before



Box_Cox on all Data



If we examine the Q–Q plot for the model trained on the *entire* dataset, we immediately see why this model is invalid. The upper tail of the plot sharply deviates upward after approximately the second theoretical quantile. This spike corresponds to all observations with non-zero claim amounts—i.e., the policyholders who actually experienced a crash. Because 75% of the data consists of zeros, the model is essentially trying to fit two fundamentally different distributions simultaneously: a large mass at zero and a long, continuous right tail for crash costs. The resulting Q–Q pattern shows that the linear model cannot capture this mixture distribution, confirming that a full-dataset severity model is statistically mis-specified and inappropriate for predicting `TARGET_AMT`.

Model Selected

Based on the modeling results, I recommend using the logistic regression model fitted with the `glm()` function as the final model for predicting crash occurrence (`TARGET_FLAG`). This model demonstrated strong overall performance, competitive AUC, and clear interpretability, making it the most suitable choice for estimating accident likelihood.

However, I will not provide predictions for `TARGET_AMT` in the evaluation set. Despite extensive testing—including untransformed, log-transformed, and Box–Cox transformed linear models—I was unable to identify a severity model with sufficient explanatory power or reliable residual behavior. The available predictors do not capture key determinants of claim cost (such as injury severity, collision type, repair estimates, or environmental factors), resulting in weak or unstable models. Therefore, no regression model tested offered a robust or valid explanation of variance in crash amounts.

```
Call:
NULL

Coefficients:
                                        Estimate Std. Error z value Pr(>|z|)
(Intercept)                           -2.629e+00  2.859e-01  -9.197  < 2e-16
INDEX                                  3.271e-06  9.798e-06   0.334 0.738467
KIDSDRIV                               3.933e-01  6.125e-02   6.421 1.36e-10
AGE                                   -1.347e-03  4.017e-03  -0.335 0.737329
HOMEKIDS                               4.577e-02  3.693e-02   1.239 0.215241
YOJ                                   -9.041e-03  7.041e-03  -1.284 0.199120
INCOME                                -3.653e-06  1.076e-06  -3.393 0.000691
PARENT1Yes                             3.702e-01  1.097e-01   3.375 0.000737
HOME_VAL                              -1.065e-06  3.174e-07  -3.356 0.000791
MSTATUSYes                            -5.288e-01  8.163e-02  -6.479 9.26e-11
SEXM                                   8.399e-02  1.121e-01   0.749 0.453851
EDUCATIONBachelors                    -3.917e-01  1.160e-01  -3.377 0.000732
`\\`EDUCATIONHigh School\\``           1.531e-02  9.531e-02   0.161 0.872380
EDUCATIONMasters                      -3.056e-01  1.793e-01  -1.705 0.088275
EDUCATIONPhD                          -1.686e-01  2.138e-01  -0.789 0.430210
JOBClerical                            1.136e-01  1.071e-01   1.061 0.288722
JOBDoctor                             -7.471e-01  2.875e-01  -2.599 0.009361
`\\`JOBHome Maker\\``                 -5.506e-02  1.507e-01  -0.365 0.714900
JOBLawyer                             -1.821e-01  1.880e-01  -0.968 0.332898
JOBManager                            -8.577e-01  1.398e-01  -6.135 8.51e-10
JOBProfessional                       -1.424e-01  1.200e-01  -1.187 0.235179
JOBSelfEmployed                       -3.670e-01  1.903e-01  -1.929 0.053745
JOBStudent                            -6.839e-02  1.273e-01  -0.537 0.591097
JOBUnspecified                         1.498e-01  3.693e-01   0.406 0.685108
TRAVTIME                               1.469e-02  1.884e-03   7.799 6.23e-15
CAR_USEPrivate                        -7.760e-01  9.270e-02  -8.371  < 2e-16
BLUEBOOK                              -2.076e-05  5.265e-06  -3.942 8.08e-05
TIF                                   -5.547e-02  7.351e-03  -7.546 4.50e-14
`\\`CAR_TYPEPanel Truck\\``            5.714e-01  1.622e-01   3.524 0.000425
CAR_TYPEPickup                         5.568e-01  1.008e-01   5.526 3.27e-08
`\\`CAR_TYPESports Car\\``             1.022e+00  1.299e-01   7.866 3.66e-15
CAR_TYPESUV                            7.649e-01  1.113e-01   6.872 6.35e-12
CAR_TYPEVan                            6.168e-01  1.267e-01   4.867 1.13e-06
RED_CARyes                            -2.085e-02  8.661e-02  -0.241 0.809807
OLDCLAIM                              -1.397e-05  3.913e-06  -3.571 0.000355
CLM_FREQ                               1.982e-01  2.857e-02   6.936 4.02e-12
REVOKEDYes                             8.893e-01  9.134e-02   9.736  < 2e-16
```

```
MVR_PTS                                  1.122e-01  1.362e-02   8.234  < 2e-16
CAR_AGE                                 -5.032e-04  7.547e-03  -0.067 0.946840
`\\`URBANICITYHighly Urban/ Urban\\``  2.387e+00  1.129e-01  21.144  < 2e-16

(Intercept)                             ***
INDEX
KIDSDRIV                                ***
AGE
HOMEKIDS
YOJ
INCOME                                  ***
PARENT1Yes                              ***
HOME_VAL                                ***
MSTATUSYes                              ***
SEXM
EDUCATIONBachelors                      ***
`\\`EDUCATIONHigh School\\``
EDUCATIONMasters                        .
EDUCATIONPhD
JOBClerical
JOBDoctor                               **
`\\`JOBHome Maker\\``
JOBLawyer
JOBManager                              ***
JOBProfessional
JOBSelfEmployed                         .
JOBStudent
JOBUnspecified
TRAVTIME                                ***
CAR_USEPrivate                          ***
BLUEBOOK                                ***
TIF                                     ***
`\\`CAR_TYPEPanel Truck\\``             ***
CAR_TYPEPickup                          ***
`\\`CAR_TYPESports Car\\``              ***
CAR_TYPESUV                             ***
CAR_TYPEVan                             ***
RED_CARyes
OLDCLAIM                                ***
CLM_FREQ                                ***
REVOKEDYes                              ***
MVR_PTS                                 ***
CAR_AGE
```

```
`\\`URBANICITYHighly Urban/ Urban\\`` ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 9404.0  on 8154  degrees of freedom
Residual deviance: 7292.3  on 8115  degrees of freedom
AIC: 7372.3

Number of Fisher Scoring iterations: 5


Confusion Matrix and Statistics

          Reference
Prediction   No  Yes
       No  5535 1254
       Yes  472  894

               Accuracy : 0.7884
                 95% CI : (0.7793, 0.7972)
    No Information Rate : 0.7366
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3823

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.4162
            Specificity : 0.9214
         Pos Pred Value : 0.6545
         Neg Pred Value : 0.8153
             Prevalence : 0.2634
         Detection Rate : 0.1096
   Detection Prevalence : 0.1675
      Balanced Accuracy : 0.6688

       'Positive' Class : Yes
```

Predictions

	TARGET_FLAG	PROBABILITY
12	Yes	0.5934129
13	Yes	0.8535058
16	Yes	0.6143612
17	Yes	0.6717466
19	Yes	0.5878506
40	Yes	0.5438705
42	Yes	0.5718318
44	Yes	0.5678644
50	Yes	0.6515241
53	Yes	0.7857003
60	Yes	0.5655090
67	Yes	0.8018129
68	Yes	0.6013007
73	Yes	0.6961675
75	Yes	0.7002078
81	Yes	0.5596722
86	Yes	0.5486320
90	Yes	0.7704456
102	Yes	0.5582588
103	Yes	0.6284113
104	Yes	0.6983618
111	Yes	0.6688092
115	Yes	0.6392941
118	Yes	0.6596362
119	Yes	0.5625891
122	Yes	0.7922069
123	Yes	0.6414239
137	Yes	0.7856060
138	Yes	0.5602126
142	Yes	0.7457301
146	Yes	0.5536438
151	Yes	0.6024202
153	Yes	0.7782082
160	Yes	0.5009801
165	Yes	0.6891534
172	Yes	0.5820545
174	Yes	0.8168194
177	Yes	0.5220745
178	Yes	0.6035907

179	Yes	0.7639490
180	Yes	0.6900857
181	Yes	0.5635962
191	Yes	0.5198901
192	Yes	0.7198615
196	Yes	0.5353803
207	Yes	0.8208924
213	Yes	0.6356019
227	Yes	0.6284536
240	Yes	0.6782173
243	Yes	0.6119718
250	Yes	0.5435674
251	Yes	0.5065193
252	Yes	0.6269000
259	Yes	0.5694168
269	Yes	0.8815471
271	Yes	0.5363851
274	Yes	0.6679381
277	Yes	0.6660390
289	Yes	0.5228512
290	Yes	0.7483168
298	Yes	0.5156490
311	Yes	0.7207193
314	Yes	0.9145867
319	Yes	0.5945918
322	Yes	0.6253466
325	Yes	0.7397864
327	Yes	0.5796888
333	Yes	0.7922792
338	Yes	0.5323476
341	Yes	0.5956877
342	Yes	0.5449244
344	Yes	0.6556079
353	Yes	0.8539518
354	Yes	0.7726979
356	Yes	0.6274185
361	Yes	0.6472501
364	Yes	0.6263996
376	Yes	0.6254485
390	Yes	0.6446164
412	Yes	0.7866588
415	Yes	0.6202803
421	Yes	0.5562226

422	Yes	0.6527011
423	Yes	0.7534879
429	Yes	0.5013553
436	Yes	0.7079558
449	Yes	0.6448923
450	Yes	0.7245858
458	Yes	0.8216495
467	Yes	0.8111636
468	Yes	0.6028882
472	Yes	0.8026790
477	Yes	0.6892681
478	Yes	0.8586337
485	Yes	0.7813656
486	Yes	0.5920019
490	Yes	0.7121224
491	Yes	0.5715303
496	Yes	0.6278379
503	Yes	0.7576291
505	Yes	0.6403445
517	Yes	0.6678576
520	Yes	0.5216708
548	Yes	0.8298988
567	Yes	0.9478280
570	Yes	0.5786254
579	Yes	0.5135409
582	Yes	0.5291750
584	Yes	0.7861065
589	Yes	0.7443204
595	Yes	0.5241923
596	Yes	0.5965388
597	Yes	0.5173122
600	Yes	0.6179290
601	Yes	0.5569555
607	Yes	0.5823971
620	Yes	0.6273441
626	Yes	0.7398475
627	Yes	0.5588081
630	Yes	0.5559643
638	Yes	0.5791521
653	Yes	0.7857355
672	Yes	0.5722556
673	Yes	0.5390025
708	Yes	0.8084010

721	Yes	0.5449540
731	Yes	0.5791216
732	Yes	0.5034009
741	Yes	0.6722510
743	Yes	0.7326385
747	Yes	0.6755034
753	Yes	0.5371586
754	Yes	0.6585268
762	Yes	0.7698573
765	Yes	0.5912801
766	Yes	0.6783603
774	Yes	0.6253691
782	Yes	0.8094033
798	Yes	0.6640043
799	Yes	0.6560630
818	Yes	0.5197157
819	Yes	0.5509555
821	Yes	0.5915601
823	Yes	0.6195457
825	Yes	0.6099745
833	Yes	0.5052761
849	Yes	0.6636393
850	Yes	0.7229323
851	Yes	0.5588775
859	Yes	0.8043803
862	Yes	0.6351870
867	Yes	0.7054491
870	Yes	0.6312149
872	Yes	0.5861545
874	Yes	0.8664922
885	Yes	0.5675002
887	Yes	0.6981789
903	Yes	0.6973173
907	Yes	0.5923436
911	Yes	0.6761749
917	Yes	0.7356086
918	Yes	0.5681439
929	Yes	0.7191665
932	Yes	0.8920306
941	Yes	0.7504280
965	Yes	0.6933782
970	Yes	0.6591119
982	Yes	0.6514010

983	Yes	0.5113052
984	Yes	0.5417888
985	Yes	0.5894832
989	Yes	0.5829772
1001	Yes	0.7053895
1002	Yes	0.5816402
1022	Yes	0.5579723
1024	Yes	0.5625881
1025	Yes	0.7260676
1042	Yes	0.5333791
1044	Yes	0.7780462
1051	Yes	0.5094907
1052	Yes	0.7381740
1058	Yes	0.5331540
1059	Yes	0.5680419
1061	Yes	0.7741281
1069	Yes	0.5590420
1073	Yes	0.7514383
1078	Yes	0.7121717
1080	Yes	0.7027036
1081	Yes	0.5798410
1084	Yes	0.5099219
1085	Yes	0.8321025
1099	Yes	0.8085855
1102	Yes	0.5061417
1110	Yes	0.6382893
1117	Yes	0.8512421
1121	Yes	0.5473601
1126	Yes	0.6824270
1132	Yes	0.7506307
1134	Yes	0.5754068
1144	Yes	0.5995834
1147	Yes	0.7739751
1149	Yes	0.5093757
1151	Yes	0.8179982
1154	Yes	0.7655674
1171	Yes	0.5496865
1172	Yes	0.5503924
1173	Yes	0.5966026
1179	Yes	0.7451154
1181	Yes	0.8247790
1184	Yes	0.8687210
1193	Yes	0.6162732

1206	Yes	0.6189228
1212	Yes	0.5811215
1216	Yes	0.5256943
1222	Yes	0.6934895
1224	Yes	0.6690168
1229	Yes	0.6104726
1233	Yes	0.6936443
1237	Yes	0.7300612
1245	Yes	0.7268973
1251	Yes	0.7584009
1256	Yes	0.7479629
1263	Yes	0.7813819
1280	Yes	0.7108129
1286	Yes	0.5605894
1290	Yes	0.6831214
1298	Yes	0.6204795
1306	Yes	0.5691598
1309	Yes	0.7458805
1310	Yes	0.5302385
1312	Yes	0.8658376
1319	Yes	0.8198182
1322	Yes	0.6962155
1325	Yes	0.5161408
1341	Yes	0.5194868
1344	Yes	0.6780109
1351	Yes	0.5621934
1362	Yes	0.6157601
1366	Yes	0.5632923
1367	Yes	0.6667406
1380	Yes	0.5164322
1381	Yes	0.7136531
1390	Yes	0.7243098
1391	Yes	0.5090347
1402	Yes	0.7524505
1403	Yes	0.5345031
1409	Yes	0.6432283
1421	Yes	0.8326303
1422	Yes	0.5013227
1425	Yes	0.8523850
1437	Yes	0.5795330
1441	Yes	0.7765247
1445	Yes	0.5830042
1456	Yes	0.7180693

1476	Yes	0.5275603
1488	Yes	0.7167959
1491	Yes	0.6919709
1494	Yes	0.7184887
1503	Yes	0.7801486
1506	Yes	0.7327015
1528	Yes	0.5325117
1529	Yes	0.6976185
1538	Yes	0.7042609
1539	Yes	0.7229060
1544	Yes	0.6209977
1553	Yes	0.6243533
1561	Yes	0.5865205
1564	Yes	0.7958045
1570	Yes	0.5558890
1591	Yes	0.7979514
1600	Yes	0.5148048
1603	Yes	0.5696002
1605	Yes	0.7986799
1609	Yes	0.8286687
1615	Yes	0.6116936
1617	Yes	0.5707322
1620	Yes	0.6290845
1622	Yes	0.6511232
1629	Yes	0.8219267
1631	Yes	0.7116818
1636	Yes	0.5170780
1649	Yes	0.6015725
1654	Yes	0.5026619
1660	Yes	0.5225421
1661	Yes	0.6185369
1662	Yes	0.5550766
1663	Yes	0.6941242
1664	Yes	0.8987755
1668	Yes	0.8332762
1670	Yes	0.6576360
1672	Yes	0.7256839
1681	Yes	0.8052334
1682	Yes	0.6587367
1684	Yes	0.5117765
1695	Yes	0.6056074
1697	Yes	0.5456456
1698	Yes	0.8278622

1706	Yes	0.6392686
1708	Yes	0.7538435
1728	Yes	0.7521317
1740	Yes	0.8868444
1753	Yes	0.7367851
1760	Yes	0.5665557
1765	Yes	0.7665235
1773	Yes	0.7986871
1776	Yes	0.5033540
1778	Yes	0.5414989
1788	Yes	0.8050420
1806	Yes	0.6750650
1807	Yes	0.6710849
1812	Yes	0.5558157
1825	Yes	0.8070719
1837	Yes	0.8638543
1839	Yes	0.7460787
1870	Yes	0.6119483
1877	Yes	0.7079988
1899	Yes	0.6324956
1909	Yes	0.7774312
1933	Yes	0.6507580
1946	Yes	0.7434402
1948	Yes	0.7305077
1952	Yes	0.5430203
1960	Yes	0.7100934
1973	Yes	0.6725281
1991	Yes	0.7720208
1992	Yes	0.6088286
1993	Yes	0.6328793
1997	Yes	0.7835046
2000	Yes	0.5872670
2002	Yes	0.6453702
2004	Yes	0.6366168
2012	Yes	0.9592798
2017	Yes	0.7771902
2018	Yes	0.5355239
2029	Yes	0.6630596
2034	Yes	0.6112743
2038	Yes	0.5542369
2047	Yes	0.5268626
2054	Yes	0.6142679
2067	Yes	0.8763177

2072	Yes	0.5407227
2076	Yes	0.6967664
2079	Yes	0.9190444
2089	Yes	0.5493558
2095	Yes	0.6823127
2098	Yes	0.5773514
2099	Yes	0.5878592
2100	Yes	0.6107748
2102	Yes	0.8819983
2106	Yes	0.6205217
2110	Yes	0.8005016
2118	Yes	0.9365366
2122	Yes	0.7265512