Technical Report on Car Insurance Claim

Introduction

This report focuses on doing logistic regression analysis because the goal is to analyze the background of policyholders who filed a car insurance claim because they were in a car accident. It can be used further to adjust the rates of car insurance for certain policyholders who are likely in car accidents and lower risk policyholders.

Methodology, Analysis and Findings

Methodology

DV(Y): CLAIM_FLAG (CLAIM_FLAG is a binary)

IV(numerical variables): KIDSDRIV AGE HOMEKIDS YOJ INCOME HOME_VAL TRAVTIME CLM_FREQ BLUEBOOK TIF OLDCLAIM MVR_PTS C AR_AGE CLM_FREQ CLM_AMT IV (qualitative variables): PARENT1 MSTATUS GENDER CAR_USE RED_CAR REVOKED URBANICITY EDUCATION (EDU) OCCUPATION CAR_TYPE (I didn't use OCCUPATION CAR_TYPE)

Step1: Filling the missing value with "."

INCOME has 570 missing values, YOJ has 548 missing values, AGE has 7 missing values, HOME_VAL has 575 missing values, and CAR_AGE has 639 missing values.

Step2: handle missing values by deleted cases that contain more than two missing values.

Delete the case when there're two variables that have missing values and the CLAIM_FLAG is 0.

Step 3: handle text values in order to use in analysis for dummy variables.

Create an EDU column: Education ≤ high school = 1, Education = bachelors = 2, education = Master or PhD = 3. I also handle the URBANICITY column by setting "Highly Rural/ Rural" =0 and "Highly Urban/ Urban"=1.

CAR_TYPE is various, so I eliminate it. I decided not to use and clean OCCUPATION because it's related to income. Next, Birth is the same as Age, so it's excluded. Step4: import data into SAS to get frequency.

Step4: import data into SAS.

Use excel to check the columns of numerical variables are general because they can't be read in SAS. Run the Descriptives to check numerical independent variables in SAS to get mean, min, median, max, and then I can use the mean to substitute. Using CLM_AMT to run a histogram, and I found out it's not normal. The median is 0. I also run frequency to check independent qualitative variables (PARENT1 MSTATUS GENDER CAR_USE RED_CAR REVOKED URBANICITY EDUCATION CLAIM_FLAG). Appendix E.1and E.1.2

Step5: fill the missing values with mean, mode, median, or similar cases.

Use Birth to calculate age, and people who were born in November and December will be one year younger than people who were born in the same year. Therefore, I filled in the missing values, and there's no missing value in AGE.

USE 50TH pctl to fill the missing data of home_val, and use 8 years to fill the missing car years. The missing values of YOJ use INCOME to determine should replace with mean, median, or similar case. I fill the missing YOJ with 0 when the income is $0. I use mean 10 years to fill some missing value, and the mean of people who are the same age, using similar cases like age, education, and occupation. After using excel to handle missing values, there're 10231 cases left.

Step6: Data exploration stage – I use boxplots and frequency tables to compare the difference of each predictors when CLAIM_FLAG= 1 and CLAIM_FLAG=0.

Step7: I run a full model and do diagnostics to check outliers and multicollinearity. Step8: I remove non-significant variables by standardized coefficients. Total cases are 10230. Step9: Random select 75% of data into training set (7673 cases) and 25% of data into test set (2557 cases).

Step10: Use stepwise, forward, backward to do model selection to analyze my training set.

Step11: Select the best model by comparing performance between the test and training set.

Analysis, result, finding

In SAS, I created dummy variables for GENDER, EDU, PARENT1, MSTATUS, CAR_USE, and RED_CAR, REVOKED. In general, I set "Yes" = 1 and "No" =0, Male is 1 and Female is 0. Deduh, dedub and dedum for EDU because I want to use all three levels, dgender for GENDER, dparent1 for PARENT1, dmstatus for MSTATUS, dcarused for CAR_USE (set "Commercial"=1), dredcar for RED_CAR, drevoked for REVOKED.

I check correlation for numerical variables and it doesn't show any multicollinearity between them. (KIDSDRIV AGE HOMEKIDS YOJ INCOME HOME_VAL TRAVTIME BLUEBOOK TIF OLDCLAIM MVR_PTS CAR_AGE CLM_FREQ) Appendix E.2

Using logistic regression to run a standardized coefficient of the model by every independent variable, and this table can be used for excluding the non-significant variables because the variables will be removed if the values of Pr> ChiSq is not <0.05.

After seeing the results including every independent variables, I chose to remove non-significant predictors one at a time. Running a model again also diagnostics, outliers, and influential points. If |Dfbeta|>0.02, it's an influential point. I use 2/sqrt (10231). I remove one outlier in the model because it's more than ±3 in Pearson Residual and Deviance Residual. There're 23 predictors in the model, which is too much, therefore, I decided to remove more non-significant variables before I split the data into training set and test set.

The first time approach, I used 15 predictors to do model selection in the training set, CLM_AMT, KIDSDRIV, HOMEKIDS, YOJ, HOME_VAL ,BLUEBOOK, CAR_AGE, CLM_FREQ, dedub, dedum, dgender, dcarused, dredcar, drevoked, URBANICITY. In model selection, I used stepwise, forward, backward methods. I got three predictors, CLM_AMT, KIDSDRIV, and CLM_FREQ in my final model. Three methods have the same AIC, SC, LIKELIHOOD ratio, and r-square. After I see the result in the test set, I realized CLM_AMT as predictor is bias because it's a response and it only happens when CLM_FLAG=1. The TP and TP are 100% correct, and FN and FP are 0. Therefore, CLM_AMT can't be the predictors in the model that dependent variable is CLM_FLAG. (Appendix E.3) The second time I approach, I remove CLM_AMT from my model. I do the diagnostic again by using a frequency table and see if people who have a high school degree or less are likely to have car accidents. (Appendix E.4) My predictors are KIDSDRIV AGE INCOME HOMEKIDS YOJ HOME_VAL BLUEBOOK TIF MVR_PTS OLDCLAIM CAR_AGE CLM_FREQ deduh dedub dedum dgender dparent1 dcarused dredcar drevoked URBANICITY. (Appendix E.5)The standardized coefficients show AIC is 11887.869, SC is 11895.103, Likelihood Ratio is 2427.6084, and Pr > ChiSq is <.0001. I need to remove variables, AGE, dredcar, dedub, and dedum because dedub has no output. Pr> ChiSq of Age, deredcar, and dedum is more than 0.8.

Next, I used 18 predictors to do my model selections stepwise, backward, and forward motheds. The predictors are KIDSDRIV HOMEKIDS YOJ INCOME HOME_VAL TRAVTIME BLUEBOOK TIF MVR_PTS CAR_AGE CLM_FREQ deduh dgender dparent1 dmstatus dcarused drevoked URBANICITY. The results of these three methods are the same. AIC is 8935.314, SC is 8942.259, R Square is 0.2180, Likelihood Ratio is 1886.6573. I have 17 predictors that are significant in my model, but the odds ratio estimates chart shows HOME_VAL, INCOME and BLUEBOOK equal to 1 in my models, so I decided to remove one of them each time, and 17 predictors are too many in the final model. Therefore, I did at least four models by removing different predictors and to compare their AIC, SC, R Square, and likelihood ratio. The first model has 15 predictors because I removed INCOME, HOME_VAL, and I got R-Square is 0.2096, Likelihood Ratio is 1804.3380. After that, I found out by removing the bluebook the likelihood ratio will be decreased. The R-square is decreased. The AIC and SC remains the same.

These two models are very similar and I decided to compare their test performance. $Ho:\beta1=\beta2=0$ $Ha:$all $\beta i\neq0$ They both reject Ho.

Final model (3) – (Removed homeval income homekids) 14 predictors

R-Square 0.2088 Max-rescaled R-Square 0.3036| Likelihood Ratio 1797.3030

Threshold 0.3	Threshold 0.35	Threshold 0.2
sensitivity 0.692878 **accuracy 0.742667** precision 0.508715 specificity 0.760489	sensitivity 0.633531 accuracy 0.768088 precision 0.552393 specificity 0.816251	sensitivity 0.830861 accuracy 0.650763 precision 0.418223 specificity 0.586298

Final model (4)- (REMOVE INCOME HOME_VAL HOMEKIDS BLUEBOOK) 13 predictors

Appendix E.6

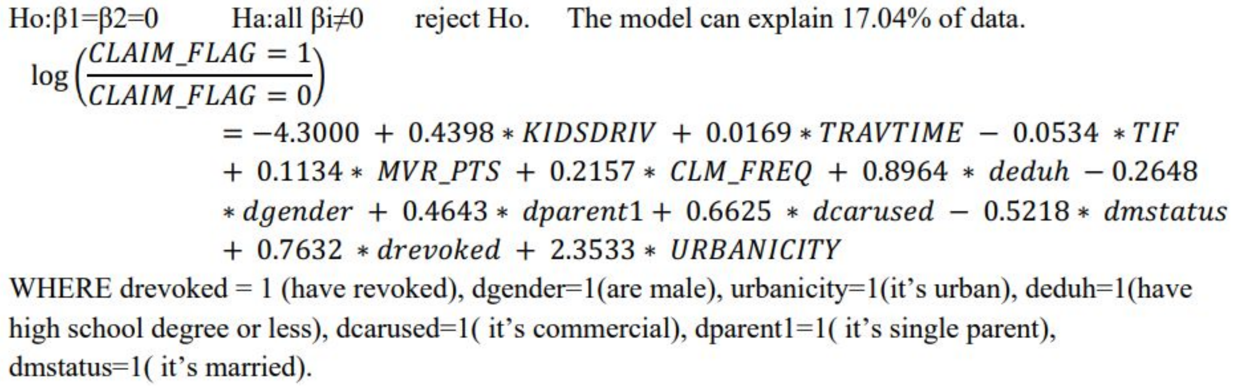
R-Square 0.1991 Max-rescaled R-Square 0.2895|Likelihood Ratio 1703.9313

Threshold 0.3	Threshold 0.35	Threshold 0.2
sensitivity 0.691395 accuracy 0.735628 precision 0.498929 specificity 0.75146	sensitivity 0.597923 accuracy 0.755964 precision 0.533069 specificity 0.812533	**sensitivity 0.847181** **accuracy 0.654673** precision 0.42265 specificity 0.585767

I want to have an overall great accuracy and sensitivity, so I chose final model 4 as my best model because the accuracy is 73.56% and sensitivity is 69.14% even though its likelihood ratio is slightly lower but it only has 13 predictors.

There are 13 predictors in my final model to test my test set.

KIDSDRIV TRAVTIME BLUEBOOK TIF MVR_PTS CLM_FREQ deduh dgender dparent1 dmstatus dcarused drevoked URBANICITY



Ho:β1=β2=0 Ha:all βi≠0 reject Ho. The model can explain 17.04% of data.

$$\log\left(\frac{CLAIM\_FLAG = 1}{CLAIM\_FLAG = 0}\right)$$
$$= -4.3000 + 0.4398 * KIDSDRIV + 0.0169 * TRAVTIME - 0.0534 * TIF + 0.1134 * MVR\_PTS + 0.2157 * CLM\_FREQ + 0.8964 * deduh - 0.2648 * dgender + 0.4643 * dparent1 + 0.6625 * dcarused - 0.5218 * dmstatus + 0.7632 * drevoked + 2.3533 * URBANICITY$$

WHERE drevoked = 1 (have revoked), dgender=1(are male), urbanicity=1(it's urban), deduh=1(have high school degree or less), dcarused=1(it's commercial), dparent1=1(it's single parent), dmstatus=1(it's married).

The conclusion is increase one driving kid, the car was in a crash will be increased by 55.2%, the travel time increases 1 the chance of the car be in a car accident will be increased by 1.7%, TIF (time in force) increases 1 the chance of the car be in a car accident will be decreased by 5.2%, increase one motor vehicle record point (MVR_PTS) the chance of the car be in a car accident will be increased by 12%, increase one claim(CLM_FREQ) the chance of the car be in a car accident will be increased by 24.1%, policyholder is high school degree or less the chance of the car be in a car accident will be increased by 1145.1%. When the gender is male (dgender) the chance of the car being in a car accident will be decreased by 23.3%. When a policyholder is a single parent, the chance of the car being in a car crash will be increased by 59.1%. When the car used is for commercial (dcarused), the chance of the car being in a car crash will be increased by 9.4%. When the policyholder is married (dmstatus), the chance of the car being in a car crash will be decreased by 40.7%. When the policyholder's license has been revoked (drevoked), the chance of the car being in a car crash will be increased by 1145%. When a policyholder is in urban (URBANICITY), the chance of the car being in a crash will be increased by 9520%. In conclusion, a policyholder is in urban, has a high school degree, the license has been revoked will have a higher chance to be in a car crash, so they will be in the high risk group.

Future Work

The model can be used to predict the likelihood of a car claim because the policyholder is in a car crash by checking the policyholder's background. The future work can do more on grouping policyholders and creating a risk level chart, and then using it to adjust insurance rates or create new insurance products.

References

- Allison, P.(2012) Logistic Regression Using SAS: Theory and Application, Second Edition. *SAS Institute*
- Li, Arthur. (2013) A Tutorial on PROC LOGISTIC https://www.mwsug.org/proceedings/2013/RX/MWSUG-2013-RX08.pdf Accessed March 15, 2019