UNITED STATES MILITARY ACADEMY

MA478: GENERALIZED LINEAR MODELS

HOMEWORK 2

MA478: APPLIED STATISTICS

SECTION H2

COL CLARK, NICHOLAS

By

CADET JOSHUA BLACKMON '24, CO C1

WEST POINT, NEW YORK

15 FEBRUARY 2024

BLACKMON.JOSHUA.LEE.1381780356 Digitally signed by BLACKMON.JOSHUA.LEE.1381780356 Date: 2024.02.15 00:31:00 -05'00' MY DOCUMENTATION IDENTIFIES ALL SOURCES USED AND ASSISTANCE RECEIVED IN COMPLETING THIS ASSIGNMENT.

_____ I DID NOT USE ANY SOURCES OR ASSISTANCE REQUIRING DOCUMENTATION IN COMPLETING THIS ASSIGNMENT.

SIGNATURE: BLACKMON.JOSHUA.LEE.1381780356 Digitally signed by BLACKMON.JOSHUA.LEE.1381780356 Date: 2024.02.15 00:31:23 -05'00'

MA478: HW2

CDT J. Blackmon

Abstract

The goal of this paper is to adequately predict whether or not someone will crash and the associated cost with each crash through both linear and logistic regression. This is focused on within the context of an insurance company's usage so the data associated with the crash are as expected. After cleaning the data by reworking the categories, blending the missing values with means, and accounting for the collinearity within the data, we settled on using stepwise regression based on AIC for the first model and the second based on BIC. With this, we have two models that can predict both who crashes in a vehicle and how much it costs based on the data given.

Any abstract should concisely describe the problem, method, and implications of the analysis that follows.

1 Introduction

Predicting car crashes and assessing the consequential costs is critical for public safety and economic planning. Furthermore, understanding the economic ramifications stemming from crashes, medical expenses, property damage, and lost productivity, underscores the urgency of predictive analytics in this domain especially for determining insurance prices. This paper delves into exploring linear and logistic regression for car crash occurrences and the comprehensive assessment of associated costs, aiming to contribute to enhanced road safety strategies and informed decision-making processes.

2 Data Preparation

The focus is to predict the probability that a person will get into an accident and the cost associated with each crash. Data given to achieve this can be shown in Figure 1. With these accumulating into the features used, we focus on predicting the Target Flag for the outcome of a crash and the Target Amount for the amount cut. To start, we cleaned up the data where typos were corrected and missing values were corrected. All missing values are now replaced with the row average, ensuring that while the data remains complete, it also maintains the integrity of each row without skewing its specific characteristics. This replacement strategy applies universally across all categories, except for the "Job Category" column, where missing values are uniformly designated as "Other".

3 Data Exploration

Moving forward, we examine any features that display correlation. After sorting through all quantitative variables, we find that car value and income are correlated with correlation coefficients 0.543. With this present, correlated features may not provide additional information gain and could lead

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 then 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

Figure 1: Data Table with definitions and descriptions.

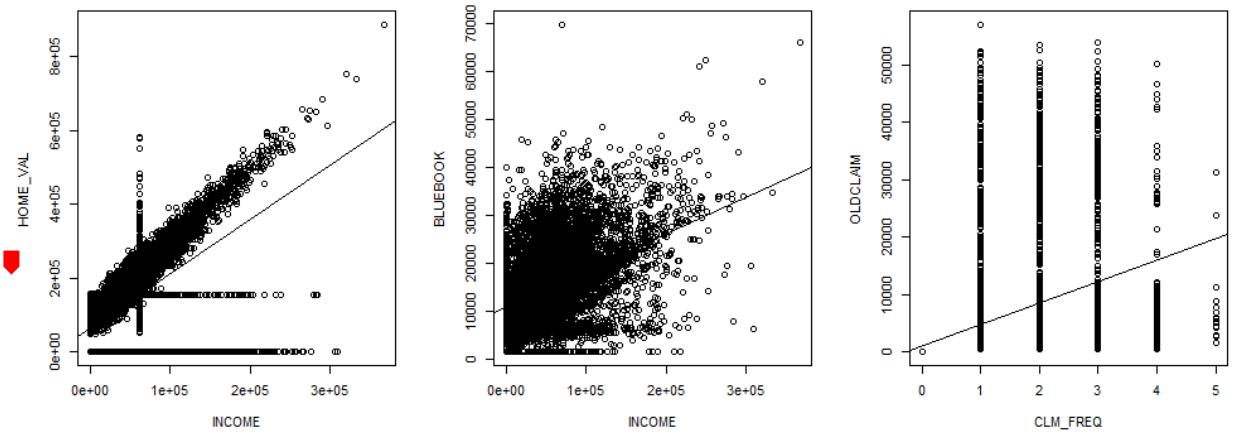

Figure 2: Plots representing the 3 greatest showcases of multicollinearity in the data.

to overfitting through multicollinearity. This can cause issues in the interpretation of coefficients and lead to unstable estimates. To avoid this, we will remove one of the variables in future use to avoid redundancy. There are other examples of less severe multicollinearity surrounding the history of claims and home value but we will adjust for it in the model creation. These can be seen in Figure 2.

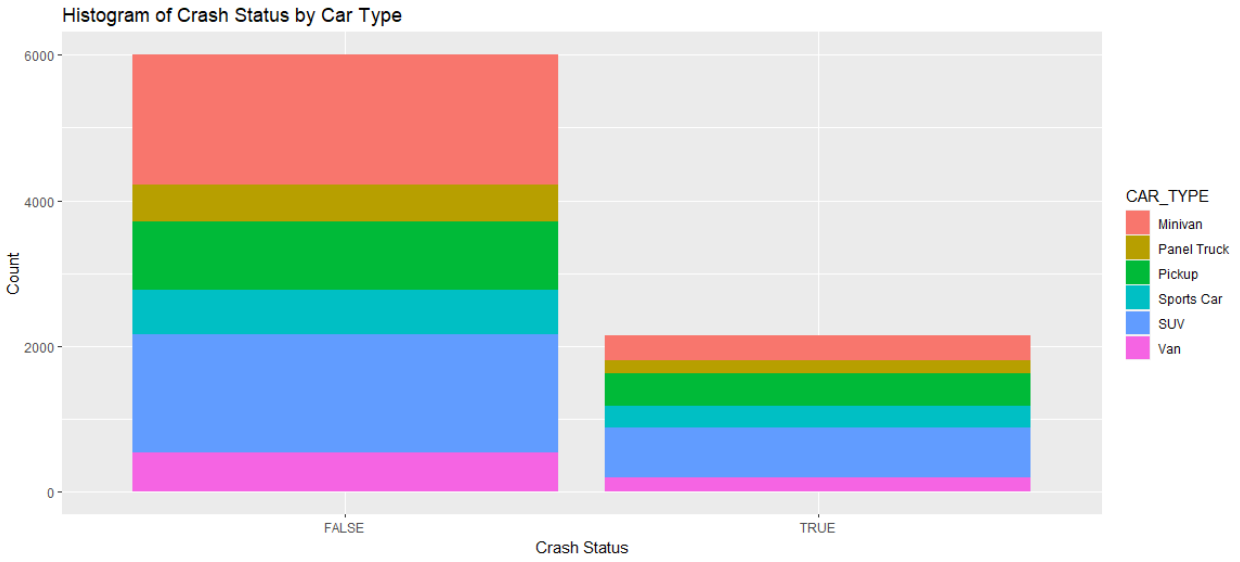

Figure 3: Histogram showcasing the Crash Status by Car Type

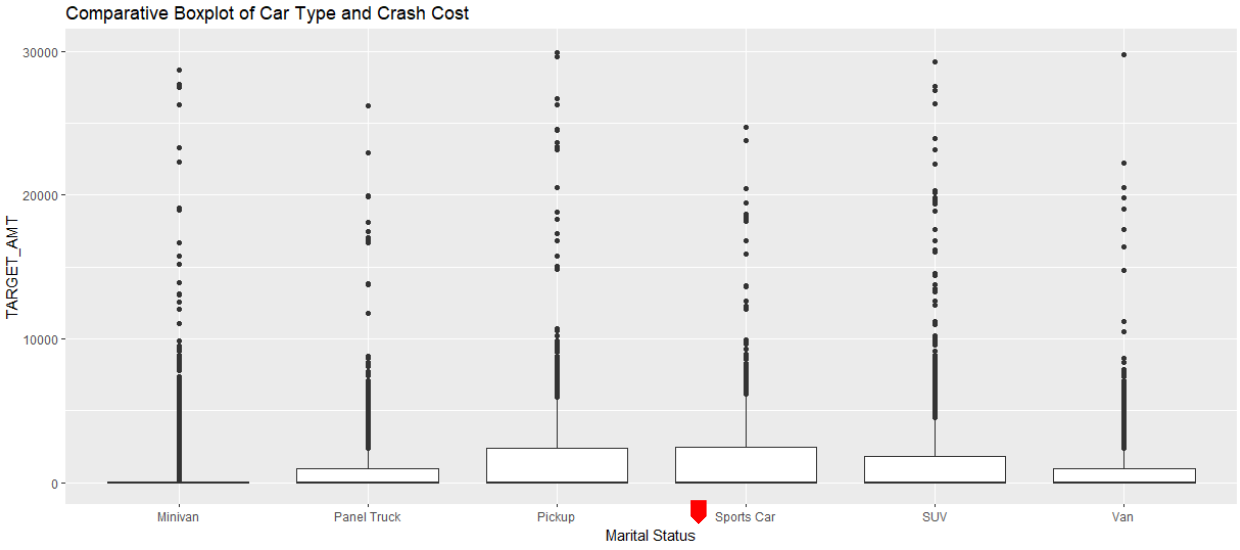

Figure 4: Boxplot showcasing the Crash Amount by Car Type

In this context, we examine potential overarching patterns. It becomes evident that these patterns underscore how the choice of vehicle by a customer correlates with distinct variations in collision occurrences. Figure 3 illustrates that while the proportion of minivans and SUVs involved in accidents is higher, a larger percentage of pickup trucks and sports cars have been in crashes. Specifically, pickup trucks and sports cars account for 20.6% and 14.1% of the total crashed vehicles, respectively.

This can be seen as well as we group the amount associated with each crash with car type. Figure 4 shows how the deviation and spread vary drastically depending on the car type. This will be important for our view into model creation.

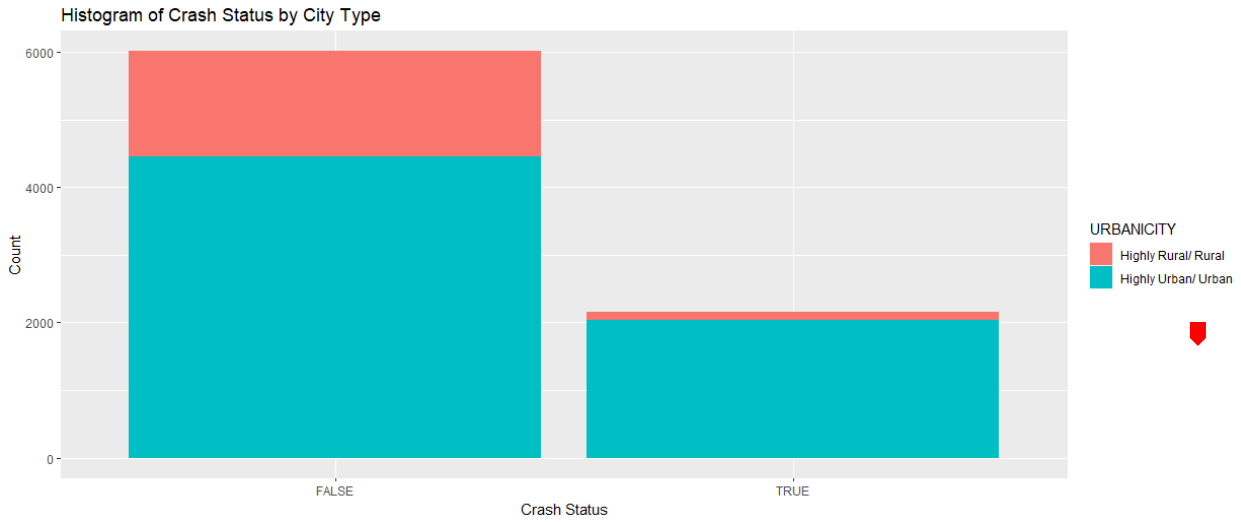

Figure 5: Histogram comparing Crash Status with City type.

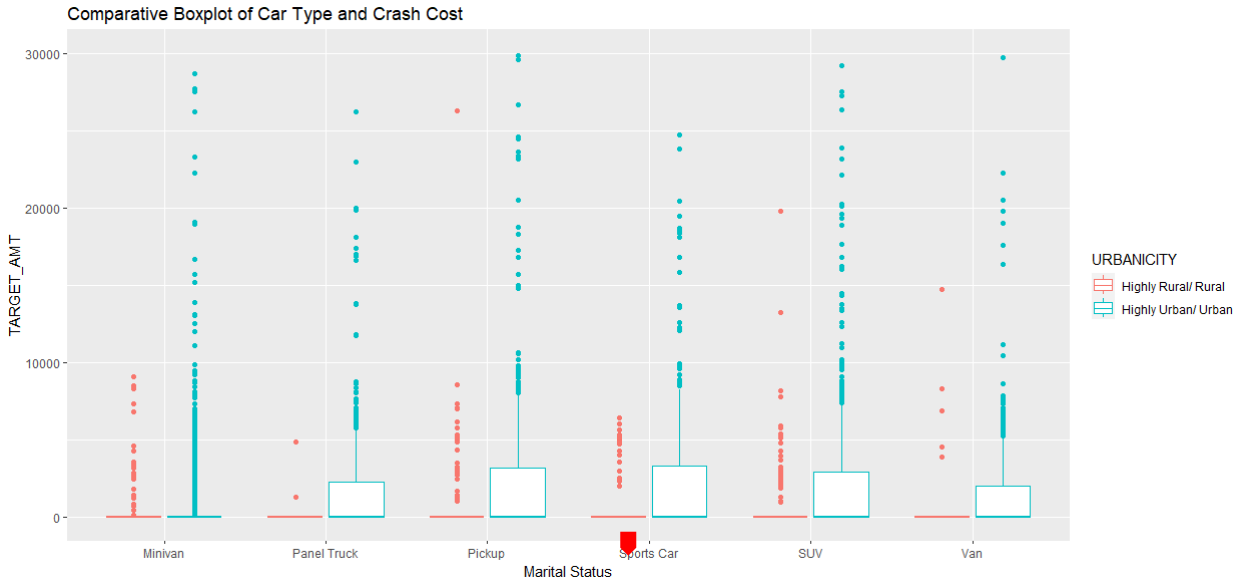

Figure 6: Boxplot comparing Crash Cost with City Type.

The same can be seen in the extreme in Work Area. In Figure 5, one can see how a strong 94.7% of crashes occur in the Urban environment. This is significant for model creation as this denotes a relationship to crash status. The same can be said for crash cost as Figure 6 shows the wildly different impact that location has on the cost of the crash.

4 Models

To predict the crash status, we used three logistic regression models. The first included all predictive variables. We prefer more simple models as they are easier to scale and deploy in production

Metric	Value
Accuracy	0.7806
Classification Error Rate	0.2194
Precision	0.6313
Sensitivity	0.4042
Specificity	0.9267
F1 Score	0.4929
AUC	0.7969

Table 1: Table showcasing the performance of the chosen model for Crash prediction.

	Predicted	
Actual	False	True
False	1113	88
True	237	193

Table 2: Table showcasing the confusion matrix for the Crash prediction model.

environments while being more interpretable and generalizable. They require fewer computational resources and are less complex to implement, making them more suitable for integration into software systems or automated processes. Additionally, simple models are easier to interpret and understand and are less prone to overfitting. For this reasoning, we move toward stepwise regression to select the most important variables (predictors) for inclusion within the model based on the Akaike information criterion (AIC) as it balances the goodness of fit of the model with its complexity. Finally, our last model identifies the best subset of predictor variables for predicting the crash status based on the Bayesian Information Criterion (BIC). The selected predictors contain the names of the predictors selected by the best model.

The best model using an AIC comparison is our second model which had an AIC of 3647.730. Model 2 as shown has its associated information with the model in Table 1.

$$\begin{aligned}\text{Model 2: glm(formula} = \text{TARGET\_FLAG} \sim\ & \text{KIDSDRIV} + \text{PARENT1} + \text{HOME\_VAL} + \\ & \text{MSTATUS} + \text{SEX} + \text{EDUCATION} + \text{JOB} + \text{TRAVTIME} + \\ & \text{CAR\_USE} + \text{BLUEBOOK} + \text{TIF} + \text{CAR\_TYPE} + \text{RED\_CAR} + \\ & \text{OLDCLAIM} + \text{CLM\_FREQ} + \text{REVOKED} + \text{MVR\_PTS} + \\ & \text{URBANICITY, family} = \text{"binomial", data} = \text{crash)}\end{aligned} \tag{1}$$

The performance of the model can be seen in the confusion matrix in Table 2 as well.

The same methodology can be seen in the choice of my three models attempting to predict crash cost. The first model is the full linear model, the second is a linear model that uses stepwise functions and the last selects the best model based on BIC. The best model this time however this time is Model 3 as shown according the the F statistic. In Table 3, the associated information is shown.

Mean Squared Error	21498041
R-squared	0.05713
F-statistic	16.418

Table 3: Table showcasing the performance of the chosen model for Crash costs.

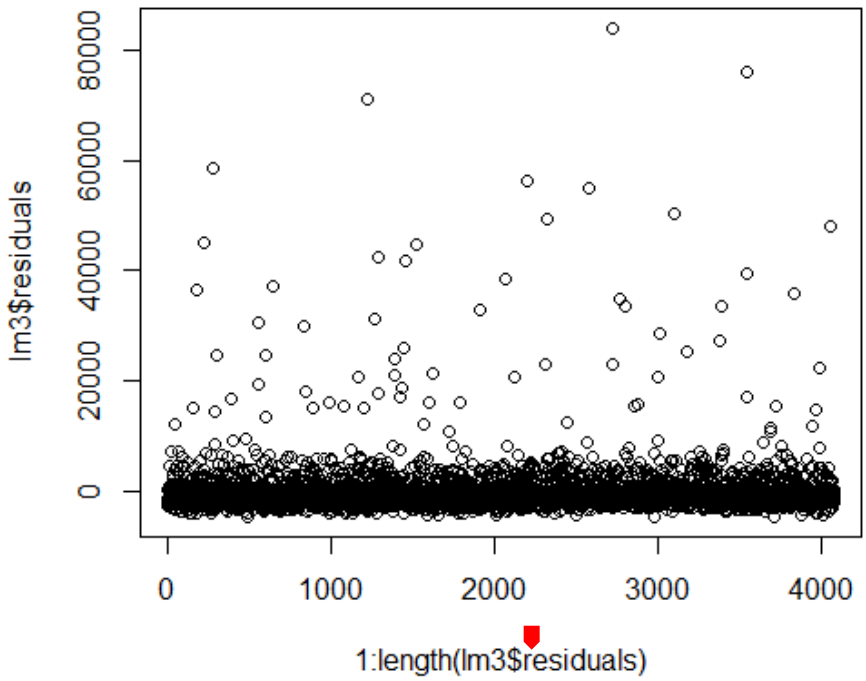

Figure 7: Scatter plot of the residuals vs predicted for independence verification.

$$\text{Model 3: TARGET\_AMT} = \beta_0 + \beta_1\text{INCOME} + \beta_2\text{PARENT1} + \beta_3\text{MSTATUS} + \beta_4\text{JOB} + \beta_5\text{CAR\_USE} + \beta_6\text{TIF} + \beta_7\text{MVR\_PTS} + \beta_8\text{URBANICITY} \quad (2)$$

Finally, we finalized checking for Linearity, Equal Variance, Normality, and Independence through the residuals. For Linearity and Equal Variance, it becomes clear there aren't any specific patterns in Figure 7. The same can be said about Figure 8 with independence.

With no specific patterns in the data, it becomes apparent that the data doesn't fail any of our validity tests.