Data 621
Tyler Baker and Jay Lee
Assignment 4

Objective

In this homework assignment, you will explore, analyze and model a data set containing approximately 8000 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.

Below is a description of each variable given in the dataset:

- TARGET_FLAG Was Car in a crash? 1=YES 0=NO
- TARGET_AMT If car was in a crash, what was the cost
- 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

Data Exploration

First of all, we are loading both csv files into R studio. It is very important to review and study the data beforehand instead of just loading it. Understanding the data is the foundation of any reports. It is very difficult to provide useful insight if you do not

even understand how the data works. We often need another professional to explain the data when it is needed. In fact, good data set is not easy to find, so checking the data is a must.

After loading the data, we can see evaluation data has 2141 observations and 26 variables, and training has 8161 observations and 26 variables. The first question in my mind is how we handle the index. Should we keep the index? Will the index be the important element in between these 2 table? After careful consideration, we decided to keep the index under back up table in case we need it to joint both table although it said that do not use it. We will use mbtraining table for the rest of the project. Deleting not useful data is something people always forget or miss. It is helpful for audience to understand your data by taking away the useless data to create a nice and clean data set.

After creating the table, we would like to explore the data to gain more knowledge on it as well as learn what limitations we may face in the later part of the project! We want to check what type the variables are. Sometimes the wrong type of data creates issues, for example you do not want the money in type of character. You want it to be numeric or integer. If we do not check and explore it at first, it may delay the project timeline. Luckily, our variable were in the correct data type.

Data Visualizations

Instead of doing data visualizations now, we will save them for later to help us understand our models. We just check the density.

Data Handling

We did have some NAs in our data. We decided that we would eliminate any columns that possessed more than 25% NAs. Luckily, none of our columns needed to be deleted. We then decided to replace the NAs with the mean average of their respective column’s mean average.

We also had some data given in the form of $x,xxx . Since, we were dealing with numeric data, we removed the dollar sign and the comma from each.

We also check the relationship between the NAs, it does not have much connection which is good.

Model Building

For this project we had to create two types of models. First we needed to model if there will be a car crash or not. This required a logical regression or a probit regression. After making both, the logical regression slightly outperformed the probit regression.

Next we had to model the cost of a car crash. This is where we ran into some trouble. We decided to make a linear regression model. However, our model had a lot of problems. It was not very trustworthy. We think that a non-linear model would fit the data better.