DATA621-Homework4-SmoothOperators

Rob Hodde, Matt Farris, Jeffrey Burmood, Bin Lin

4/10/2017

Problem Description

The 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.

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.

Using the training data set, evaluate the multiple linear regression model based on (a) mean squared error, (b) R2, (c) F-statistic, and (d) residual plots. For the binary logistic regression model, will use a metric such as log likelihood, AIC, ROC curve, etc.? Using the training data set, evaluate the binary logistic regression model based on (a) accuracy, (b) classification error rate, (c) precision, (d) sensitivity, (e) specificity, (f) F1 score, (g) AUC, and (h) confusion matrix. Make predictions using the evaluation data set.

Approach Steps:

1) Build a logistic regression model based on the TARGET_FLAG response variable.
2) Generate TARGET_FLAG predictions using the logistic regression model.
3) Build a linear regression model based on the non-zero values of the TARGET_AMT response variable.
4) Generate TARGET_AMT predictions using the linear regression model based on the non-zero values of the predicted TARGET_FLAG variable.

Data Exploration

Data Exploration

VARIABLE NAME	DEFINITION
INDEX	Identification Variable (do not use)
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
BLUEBOOK	Value of Vehicle
CAR_AGE	Vehicle Age
CAR_TYPE	Type of Car
CAR_USE	Vehicle Use
CLM_FREQ	# Claims (Past 5 Years)
EDUCATION	Max Education Level
HOMEKIDS	# Children at Home
HOME_VAL	Home Value
INCOME	Income
JOB	Job Category
KIDSDRIV	# Driving Children
MSTATUS	Marital Status
MVR_PTS	Motor Vehicle Record Points
OLDCLAIM	Total Claims (Past 5 Years)
PARENT1	Single Parent
RED_CAR	A Red Car
REVOKED	License Revoked (Past 7 Years)
SEX	Gender
TIF	Time in Force
TRAVTIME	Distance to Work
URBANICITY	Home/Work Area
YOJ	Years on Job

Below is a summary of each predictor variable's basic statistics, followed by boxplots which illustrate the spread and outliers for each variable.

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ
Min. :0.0000	Min. : 0	Min. :0.0000	Min. :16.00	Min. :0.0000	Min. : 0.0
1st Qu.:0.0000	1st Qu.: 0	1st Qu.:0.0000	1st Qu.:39.00	1st Qu.:0.0000	1st Qu.: 9.0
Median :0.0000	Median : 0	Median :0.0000	Median :45.00	Median :0.0000	Median :11.0

TARGET_FLAG	TARGET_AMT	KIDSDRIV	AGE	HOMEKIDS	YOJ
Mean :0.2638	Mean : 1504	Mean :0.1711	Mean :44.79	Mean :0.7212	Mean :10.5
3rd Qu.:1.0000	3rd Qu.: 1036	3rd Qu.:0.0000	3rd Qu.:51.00	3rd Qu.:1.0000	3rd Qu.:13.0
Max. :1.0000	Max. :107586	Max. :4.0000	Max. :81.00	Max. :5.0000	Max. :23.0
NA	NA	NA	NA's :6	NA	NA's :454

INCOME	PARENT1	HOME_VAL	MSTATUS	SEX	EDUCATION
$0 : 615	No :7084	$0 :2294	Yes :4894	M :3786	<High School :1203
: 445	Yes:1077	: 464	z_No:3267	z_F:4375	Bachelors :2242
$26,840 : 4	NA	$111,129: 3	NA	NA	Masters :1658
$48,509 : 4	NA	$115,249: 3	NA	NA	PhD : 728
$61,790 : 4	NA	$123,109: 3	NA	NA	z_High School:2330
$107,375: 3	NA	$153,061: 3	NA	NA	NA
(Other) :7086	NA	(Other) :5391	NA	NA	NA

JOB	TRAVTIME	CAR_USE	BLUEBOOK	TIF	CAR_TYPE
z_Blue Collar:1825	Min. : 5.00	Commercial:3029	$1,500 : 157	Min. : 1.000	Minivan :2145
Clerical :1271	1st Qu.: 22.00	Private :5132	$6,000 : 34	1st Qu.: 1.000	Panel Truck: 676
Professional :1117	Median : 33.00	NA	$5,800 : 33	Median : 4.000	Pickup :1389
Manager : 988	Mean : 33.49	NA	$6,200 : 33	Mean : 5.351	Sports Car : 907
Lawyer : 835	3rd Qu.: 44.00	NA	$6,400 : 31	3rd Qu.: 7.000	Van : 750
Student : 712	Max. :142.00	NA	$5,900 : 30	Max. :25.000	z_SUV :2294
(Other) :1413	NA	NA	(Other):7843	NA	NA

RED_CAR	OLDCLAIM	CLM_FREQ	REVOKED	MVR_PTS	CAR_AGE	URBANICITY
no :5783	$0 :5009	Min. :0.0000	No :7161	Min. : 0.000	Min. :-3.000	Highly Urban/ Urban
yes:2378	$1,310 : 4	1st Qu.:0.0000	Yes:1000	1st Qu.: 0.000	1st Qu.: 1.000	z_Highly Rural/ Rural
NA	$1,391 : 4	Median :0.0000	NA	Median : 1.000	Median : 8.000	NA
NA	$4,263 : 4	Mean :0.7986	NA	Mean : 1.696	Mean : 8.328	NA
NA	$1,105 : 3	3rd Qu.:2.0000	NA	3rd Qu.: 3.000	3rd Qu.:12.000	NA
NA	$1,332 : 3	Max. :5.0000	NA	Max. :13.000	Max. :28.000	NA
NA	(Other):3134	NA	NA	NA	NA's :510	NA

Based on an analysis of the box plots, the following variables have some outliers that may, or may not, exert influence on the regression results: - zn, rm, dis, black, lstat, medv

We'll next look at these variables more closely, starting with their histograms and frequency counts to better understand the nature of their distribution.

According to the description, the variables *zn*, *indus*, and *age* are area, or land, proportions. According to the statistical summary, the values for these variables are all within the range [1,100] that we would expect.

Based on our detailed review of the variables that contained outliers, the following variables could be problematic:

The predictor variable *zn* is highly right skewed, we can confirm this by comparing the median and mean where the median is 0.0, but the mean is 11.58. The frequency count plot shows how poor the distribution is due to clustering of the data at one extreme.

The predictor variable *black* is highly left skewed. We can confirm this by comparing the median and mean where the median is 391.34 and the mean is 357.12. The frequency count plot shows how poor the distribution is due to clustering of the data at one extreme.

The predictor variable *dis* is slightly right skewed. We can confirm this by comparing the median and mean where the median is 3.191 and the mean is 3.796.

Fortunately, no missing data, or NAs, were found.

The following data corrections were identified in this section:

(1) The predictor variable *chas* and the response variable *target* are categorical (binary), so we need to convert them to factors.

(2) Need to determine if there are other variables highly coorelated with the *zn* or *black* variables that do not have the severe skew and outliers. This could allow us to remove the *zn* or *black* variables from the model.

Data Preparation

Data Preparation

The variable changes we identified so far include converting the predictor variable *chas* and the response variable *target* to factors. Next we will look at how each variable correlates to all the others:

The correlation table above shows that the variable *zn* is moderately correlated to the variable *dis*. The plot of the *dis* data shows a much better distribution of values. Consequently, one possibility is to remove *zn* from the model and use *dis* instead. Before doing this, we should look at the real-world context of the two variables to determine if they are meaningfully related.

Build Models

Build Models

One method of developing multiple regression models is to take a stepwise approach. To accomplish this, we combine our knowledge from the data exploration above with logistic regression. Univariate Logistic Regression is a useful method to understand how each predictor variable interacts individually with the target

(response) variable. Looking at various statistics, we determine which variable may impact our target the most.

Here we see the selected output criteria for the linear models run with only a single predictor variable. We examine the p-value (significance), the AIC statistic (goodness-of-fit) and the AUC (Area Under Curve) to measure the potential predictive value of each variable, so we can decide whether or not to include it in our multiple regression model. We are looking for p-values below .05, AIC values as low as possible, and AUC values as high as possible.

From the above table, we can see that *chas* is the least likely to produce any meaningful inference because its p-value is well above .05 (not significant), it has the highest AIC (518, where 100 is considered excellent), and the lowest AUC (.54, where random chance would yield .50). Therefore, *chas* is the most likely candidate to be removed from our model.

Model 1

As a baseline, we start with a multiple logistic regression model that includes every predictor variable:

In this model-and in all the models- we set aside 20% of the training data and use 80% to train the model we then use the model to predict the outcome of the remaining 20% of the data. The model yields an Area Under Curve of .95, meaning it chose correctly 95% of the time.

Model 2

In this scenario we attempt to create the simplest model possible by using only one variable - the one that provides the highest overall AUC (performance) by itself. We calculate AUC for each variable separately and then select the highest result.

The best predictor variable is *nox*, yielding an AUC of .87.

Next we combine *nox* with each of the remaining variables individually and select the highest AUC result.

We find that *nox* plus *rad* is the strongest combinaton of two variables, yielding an AUC of .93.

Finally, we search for a third critical predictor by combining *nos* plus *rad* with the remaining variables, individually.

By combining three variables - *nox*, *rad* and *zn* - that is, the concentration of nitrogen oxides, access to radial highways and the proportion of land zoned for large lots, we can predict with 94% accuracy whether the crime rate at this property is above or below average. Since this is very close to the performance of the model using all variables (95%), we can be confident in using these three variables for our decision support process, and disregarding the others.

Model 3

The GLM Model summary in Model 1 illustrates the outsize impact of the predictor variable *nox* compared to all the others. It carries an Estimate of 53.3 where the next closest in magnitude is only 1.2. We thought it would be interesting to remove *nox* from the model just to see how the other variables perform without it. First we will perform a simple backward variable selection optimization process including it.

MODEL 3 WITH NOX VARIABLE The model reduces to nine variables and yields a nice low residual deviance of 133.9, compared to a null deviance of 515.3. This roughly means that the model eliminates about 80% of the error compared to choosing at random. The AUC is .947 which is roughly the same as the full model using all variables.

Let's look at what happens when we remove the *nox* variable:

MODEL 3 WITHOUT NOX VARIABLE We still have a good model - the Residual Deviance increased to 182, but that is still much better than predicting with no model at all. The AUC is now .89 - again, very good. But the AUC with only one variable *nox* was .87. And in certain trials the AUC with *nox* exceeded .95 (due to randomly selected evaluation samples).

Why is the *nox* variable so powerful? We can look back at the Correlation table for clues. More variables are significantly correlated to *nox* than any other. It is like a super-variable, somehow encapsulating the properties of the variables around it. Is it because *nox* is an indicator of so many problems, like pollution, industrial decay, lax building codes? The *nox* variable is a stellar example of a finding that opens up many paths for further research.

Below is table illustrating the various fitness parameters that describe the effectiveness of the models. All the models are good - from a practical perspective, there is no difference between them.

Choose Model

Choose Model

We like *Model 3 With Nox* the best because it eliminates some of the questionable variables - the ones with high skew and many outliers, also it eliminates the chas variable, which was shown earlier as being insignificant. Ridding the model of these variables helps provide insurance against poor decisions that could arise, *even if they do not show up in the model.*

MODEL 3 WITH NOX VARIABLE USING FULL DATASETS The Smooth Operators of R Fusion Have Struck Again.