DATA 621 - Group Assignment 3 Write Up: Logistic Regression on Crime Rates: Write Up

(Group 4) Eddie Xu, Mohamed Hassan-El Serafi, Chun Shing Leung, Keith Colella, Yina, Qiao

2024-11-03

Contents

Introduction

Crime rates in neighborhoods are an important factor in finding a new home or allocating resources to better improve the affected neighborhood. For this assignment, our group are going to explore, analyze and model a data set containing information on crime for various neighborhoods of a major city and later going to use binary logistic regression to predict whether the neighborhood will be at risk for high crime levels. A binary logistic regression is a statistic method that models the relation between a binary dependent variable and a set of independent variables. It is widely used to determine dichotomous outcomes.

In this assignment, our group explored the training data set for better understanding. We later prepared the data based on the insight and produced 3 different training data sets. After that, we evaluated all binary models and determined the best model that offers the best accuracy and completeness.

Data Exploration

Describe the size and the variables in the crime training data set. Consider that too much detail will cause a manager to lose interest while too little detail will make the manager consider that you aren't doing your job.

A simple exploratory data analysis will be conducted and the training data provided will be used to determine the property and value of the dataset. The `DataExplorer` package will be used to provide a full profile for the data frame.

Data Set

With the data set given, there are 466 rows, 13 columns and 6058 observations. There is no missing values or observations and all columns have continuous values. The `chas` variable is the only dummy variable out of 13 columns and is used to determine if the suburb borders the Charles River. Based on the histograms, both `rm` and `medv` variables are normally distributed, while other variables are skewed. Both `tax` and `rad` variables have very high outliers. That could be a data quality issue.

Given the box plot below, the variable `chas` has few outliers and the median is close to 0. Not only that, many variables including `dis`, `age`, `lstat`, `tax` and `rad` are skewed.

Based on the correlation plot, it shows that `rad` and `dis` have the highest positive correlation compared to other variables.

To determine if the dataset is compatible with the binary logistic regression model, the model is fitted and the VIF score analysis is conducted to check for any multicollinearity.

After the model fitting, the P-value is less than 0.05, showing that predictor variables may be significantly associated with the outcome. There are variables in the data set with moderate correlation between predictor variables. Both `rad` and `tax` variables have the highest VIF scores and it is over 5, showing that they are severely correlated with other predictor variables. Therefore, either variable need to removed and need to reevaluated before modeling the binary regression model.

With either `rad` or `tax` variables removed from the data set, the updated data set is reevaluated and with the updated VIF score, both variable score dropped below 5 and it appears that `tax` variable have a bigger impact. Therefore, the data set with `tax` removed will be selected for the model building.

Data Preparation

Describe how you have transformed the data by changing the original variables or creating new variables. If you did transform the data or create new variables, discuss why you did this.

Based on the outcome from our data exploratory analysis, some of the variables are identified to be skewed, so log transformations are needed to address the issue. Log transformations are applied on these variables: `indus`, `age`, and `dis`.

Since `tax` and `rad` variables have outliers, those outliers will be removed given that they might be a data quality issue. Since there is no additional information about these outliers, they are considered to be removed.

Build Model

Using the training data, build at least three different binary logistic regression models, using different variables (or the same variables with different transformations). You may select the variables manually, use an approach such as Forward or Step wise, use a different approach, or use a combination of techniques. Describe the techniques you used. If you manually selected a variable for inclusion into the model or exclusion into the model, indicate why this was done. Be sure to explain how you can make inferences from the model, as well as discuss other relevant model output.

With the training data explored and prepped, 4 different models were built to determine the best model. The first model is based on the original training data set and will set the baseline. The second data set is based on the log transformation on skewed variables. The third data set is based on the `tax` variable being removed. The last model will be based on a dataset with rows removed due to them being outliers on `tax` and `rad` variables.

Select Model

Decide on the criteria for selecting the best binary logistic regression model. Will you select models with slightly worse performance if it makes more sense or is more parsimonious? Discuss why you selected your model.

In order to select the best model, we decided to use the confusion matrix and ROC curve to help analyze each model's performance. The confusion matrix provides a better understanding in the model's prediction and the ROC curve visualizes how well the model distinguish each variables. Out of all binary models, the model with `tax` variable removed has the best performance. The AUC is higher than other 3 models and the confusion matrix shows that it has more true positive predictions as well.

Appendix

```
##  Data Exploration
# pull in the training data set
crime_training_data <- read.csv("https://raw.githubusercontent.com/eddiexunyc/crime_binary_logistic_reg

# pull in the test data
crime_test_data <- read.csv("https://raw.githubusercontent.com/eddiexunyc/crime_binary_logistic_regress

glimpse(crime_training_data)

introduce(crime_training_data)

# par on plots
par(mfrow = c(1, 4))
plot_intro(crime_training_data)
describeBy(crime_training_data)
plot_histogram(crime_training_data)

# boxplot on variables
crime_box_plot <- crime_training_data %>%
  gather(key, value, -target) %>%
  mutate(key = factor(key),
         target = factor(target)) %>%
  ggplot(aes(x = key, y = value)) +
  geom_boxplot(aes(fill = target)) +
  facet_wrap(~ key, scales = 'free', ncol = 4) +
  scale_fill_manual(values=c("lightblue", "pink")) +
  coord_flip() +
  theme_minimal()

# correlation plot on variables
par(mfrow = c(1,2))
crime_box_plot
corPlot(crime_training_data, upper = FALSE)

# fit a linear regression before VIF score
vif_model_all <- lm(target ~ ., data = crime_training_data)

summary(vif_model_all)
```

```r
# perform VIF
vif_value = vif(vif_model_all)
vif_value

# tax removed
crime_training_data_tax_removed <- crime_training_data %>%
  dplyr::select(-c(tax))
vif_model_tax <- lm(target ~., data = crime_training_data_tax_removed)
vif2_score <- vif(vif_model_tax)

# rad removed
crime_training_data_rad_removed <- crime_training_data %>%
  dplyr::select(-c(rad))
vif_model_rad <- lm(target ~., data = crime_training_data_rad_removed)
vif3_score <- vif(vif_model_rad)

# print score
vif2_score
vif3_score

## Data Preparation
# perform a log transformation on rad and dis variables
crime_training_data_transformed <- crime_training_data %>%
  mutate(log(crime_training_data$age + 1),
         log(crime_training_data$dis + 1),
         log(crime_training_data$lstat + 1))

# remove the skewed variables
crime_training_data_updated <- crime_training_data %>%
  filter(crime_training_data$rad != 24)

head(crime_training_data_updated)

## Build Model
# set seed
set.seed(123)

### Model 1
crime_binary_model_1 <- glm(crime_training_data, family = 'binomial', formula = target ~.)
summary(crime_binary_model_1)
plot(crime_binary_model_1)

### Model 2
crime_binary_model_2 <- glm(crime_training_data_transformed, family = 'binomial', formula = target ~.)
summary(crime_binary_model_2)
plot(crime_binary_model_2)

### Model 3
crime_binary_model_3 <- glm(crime_training_data_tax_removed, family = 'binomial', formula = target ~.)
summary(crime_binary_model_3)
plot(crime_binary_model_3)

### Model 4
```

```
crime_binary_model_4 <- glm(crime_training_data_updated, family = 'binomial', formula = target ~.)
summary(crime_binary_model_4)
plot(crime_binary_model_4)

## Select Model
### Model 1 Assessment
data_split_model_1 <- createDataPartition(y = crime_training_data$target, p = 0.8, list = FALSE)
crime_train_data_model_1 <- crime_training_data[data_split_model_1,]
crime_test_data_model_1 <- crime_training_data[-data_split_model_1,]
crime_binary_test_model_1 <- glm(crime_train_data_model_1, family = 'binomial', formula = target ~.)
crime_binary_prediction_1 <- predict(crime_binary_test_model_1, crime_test_data_model_1, type = "respons
crime_predicted_class_1 <- ifelse(crime_binary_prediction_1 > 0.5, 1, 0)
crime_confusion_matrix_1 <- confusionMatrix(data = as.factor(crime_predicted_class_1), reference = as.fa

print(crime_confusion_matrix_1)

roc(crime_test_data_model_1$target, crime_binary_prediction_1 , percent=TRUE, plot=TRUE, ci=TRUE, print

### Model 2 Assessment
data_split_model_2 <- createDataPartition(y = crime_training_data_transformed$target, p = 0.8, list = FA
crime_train_data_model_2 <- crime_training_data_transformed[data_split_model_2,]
crime_test_data_model_2 <- crime_training_data_transformed[-data_split_model_2,]
crime_binary_test_model_2 <- glm(crime_train_data_model_2, family = 'binomial', formula = target ~.)
crime_binary_prediction_2 <- predict(crime_binary_test_model_2, crime_test_data_model_2, type = "respons
crime_predicted_class_2 <- ifelse(crime_binary_prediction_2 > 0.5, 1, 0)
crime_confusion_matrix_2 <- confusionMatrix(data = as.factor(crime_predicted_class_2), reference = as.fa

print(crime_confusion_matrix_2)

roc(crime_test_data_model_2$target, crime_binary_prediction_2, percent=TRUE, plot=TRUE, ci=TRUE, print.a

### Model 3 Assessment
data_split_model_3 <- createDataPartition(y = crime_training_data_tax_removed$target, p = 0.8, list = FA
crime_train_data_model_3 <- crime_training_data_tax_removed[data_split_model_3,]
crime_test_data_model_3 <- crime_training_data_tax_removed[-data_split_model_3,]
crime_binary_test_model_3 <- glm(crime_train_data_model_3, family = 'binomial', formula = target ~.)
crime_binary_prediction_3 <- predict(crime_binary_test_model_3, crime_test_data_model_3, type = "respons
crime_predicted_class_3 <- ifelse(crime_binary_prediction_3 > 0.5, 1, 0)
crime_confusion_matrix_3 <- confusionMatrix(data = as.factor(crime_predicted_class_3), reference = as.fa

print(crime_confusion_matrix_3)

roc(crime_test_data_model_3$target, crime_binary_prediction_3, percent=TRUE, plot=TRUE, ci=TRUE, print.a

### Model 4 Assessment
data_split_model_4 <- createDataPartition(y = crime_training_data_updated$target, p = 0.8, list = FALSE)
crime_train_data_model_4 <- crime_training_data_updated[data_split_model_4,]
crime_test_data_model_4 <- crime_training_data_updated[-data_split_model_4,]
crime_binary_test_model_4 <- glm(crime_train_data_model_4, family = 'binomial', formula = target ~.)
crime_binary_prediction_4 <- predict(crime_binary_test_model_4, crime_test_data_model_4, type = "respons
crime_predicted_class_4 <- ifelse(crime_binary_prediction_4 > 0.5, 1, 0)
crime_confusion_matrix_4 <- confusionMatrix(data = as.factor(crime_predicted_class_4), reference = as.fa
```

```
print(crime_confusion_matrix_4)

roc(crime_test_data_model_4$target, crime_binary_prediction_4, percent=TRUE, plot=TRUE, ci=TRUE, print.a
```