DATA 621 - HW3

Andrew Bowen, Glen Davis, Shoshana Farber, Joshua Forster, Charles Ugiagbe

2023-10-23

Homework 3 - Logistic Regression

Overview:

In this homework assignment, you will explore, analyze and model a data set containing information on crime for various neighborhoods of a major city. Each record has a response variable indicating whether or not the crime rate is above the median crime rate (1) or not (0).

Your objective is to build a binary logistic regression model on the training data set to predict whether the neighborhood will be at risk for high crime levels. You will provide classifications and probabilities for the evaluation data set using your binary logistic regression model. You can only use the variables given to you (or, variables that you derive from the variables provided).

Below is a short description of the variables of interest in the data set:

Column	Description
`zn`	proportion of residential land zoned for large lots (over 25000 square feet) (*predictor variable*)
`indus`	proportion of non-retail business acres per suburb (*predictor variable*)
`chas`	a dummy var. for whether the suburb borders the Charles River (1) or not (0) (*predictor variable*)
`nox`	nitrogen oxides concentration (parts per 10 million) (*predictor variable*)
`rm`	average number of rooms per dwelling (*predictor variable*)
`age`	proportion of owner-occupied units built prior to 1940 (*predictor variable*)
`dis`	weighted mean of distances to five Boston employment centers (*predictor variable*)
`rad`	index of accessibility to radial highways (*predictor variable*)
`tax`	full-value property-tax rate per $10,000 (*predictor variable*)
`ptratio`	pupil-teacher ratio by town (*predictor variable*)
`lstat`	lower status of the population (percent) (*predictor variable*)
`medv`	median value of owner-occupied homes in $1000s (*predictor variable*)
`target`	**whether the crime rate is above the median crime rate (1) or not (0) (*response variable*)**

Data Loading:

Let's load in the training dataset.

```
train_df <- read.csv('https://raw.githubusercontent.com/ShanaFarber/businessAnalyticsDataMiningDATA621/
```

Data Exploration:

```
train_df |>
    glimpse()
```

```
## Rows: 466
## Columns: 13
## $ zn      <dbl> 0, 0, 0, 30, 0, 0, 0, 0, 0, 80, 22, 0, 0, 22, 0, 0, 100, 20, 0~
## $ indus   <dbl> 19.58, 19.58, 18.10, 4.93, 2.46, 8.56, 18.10, 18.10, 5.19, 3.6~
## $ chas    <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ nox     <dbl> 0.605, 0.871, 0.740, 0.428, 0.488, 0.520, 0.693, 0.693, 0.515,~
## $ rm      <dbl> 7.929, 5.403, 6.485, 6.393, 7.155, 6.781, 5.453, 4.519, 6.316,~
## $ age     <dbl> 96.2, 100.0, 100.0, 7.8, 92.2, 71.3, 100.0, 100.0, 38.1, 19.1,~
## $ dis     <dbl> 2.0459, 1.3216, 1.9784, 7.0355, 2.7006, 2.8561, 1.4896, 1.6582~
## $ rad     <int> 5, 5, 24, 6, 3, 5, 24, 24, 5, 1, 7, 5, 24, 7, 3, 3, 5, 5, 24, ~
## $ tax     <int> 403, 403, 666, 300, 193, 384, 666, 666, 224, 315, 330, 398, 66~
## $ ptratio <dbl> 14.7, 14.7, 20.2, 16.6, 17.8, 20.9, 20.2, 20.2, 20.2, 16.4, 19~
## $ lstat   <dbl> 3.70, 26.82, 18.85, 5.19, 4.82, 7.67, 30.59, 36.98, 5.68, 9.25~
## $ medv    <dbl> 50.0, 13.4, 15.4, 23.7, 37.9, 26.5, 5.0, 7.0, 22.2, 20.9, 24.8~
## $ target  <int> 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 0, 0, 1, 1, 0, 0, 0, 1, 1, 1, 0,~
```

The dataset consists of 466 observations of 13 variables. There are 12 predictor variables and one response variable (`target`).

All of the columns in the dataset are numeric, but the predictor variable `chas` is a dummy variable, as is the response variable `target`. We recode them as factors.

```
train_df <- train_df |>
    mutate(chas = as.factor(chas), target = as.factor(target))
```

Let's take a look at the summary statistics for the variables in the dataset.

```
remove <- c("vars", "trimmed", "mad")
describe <- train_df |>
    describe() |>
    select(-all_of(remove))
knitr::kable(describe, format = "simple")
```

	n	mean	sd	median	min	max	range	skew	kurtosis	
zn	466	11.5772532	23.3646511	0.00000	0.0000	100.0000	100.0000	2.1768152	3.8135765	1.082
indus	466	11.1050215	6.8458549	9.69000	0.4600	27.7400	27.2800	0.2885450	-1.2432132	0.317
chas*	466	1.0708155	0.2567920	1.00000	1.0000	2.0000	1.0000	3.3354899	9.1451313	0.011
nox	466	0.5543105	0.1166667	0.53800	0.3890	0.8710	0.4820	0.7463281	-0.0357736	0.005

	n	mean	sd	median	min	max	range	skew	kurtosis	
rm	466	6.2906738	0.7048513	6.21000	3.8630	8.7800	4.9170	0.4793202	1.5424378	0.032
age	466	68.3675966	28.3213784	77.15000	2.9000	100.0000	97.1000	-0.5777075	-1.0098814	1.311
dis	466	3.7956929	2.1069496	3.19095	1.1296	12.1265	10.9969	0.9988926	0.4719679	0.097
rad	466	9.5300429	8.6859272	5.00000	1.0000	24.0000	23.0000	1.0102788	-0.8619110	0.402
tax	466	409.5021459	167.9000887	334.50000	187.0000	711.0000	524.0000	0.6593136	-1.1480456	7.777
ptratio	466	18.3984979	2.1968447	18.90000	12.6000	22.0000	9.4000	-0.7542681	-0.4003627	0.101
lstat	466	12.6314592	7.1018907	11.35000	1.7300	37.9700	36.2400	0.9055864	0.5033688	0.328
medv	466	22.5892704	9.2396814	21.20000	5.0000	50.0000	45.0000	1.0766920	1.3737825	0.428
target*	466	1.4914163	0.5004636	1.00000	1.0000	2.0000	1.0000	0.0342293	-2.0031131	0.023

We can see the mean, median, standard deviations, ranges, etc. for each of the variables in the dataset.

Each predictor has 466 values, which matches the number of observations in our dataset, so there do not appear to be any missing values to address. Let's validate this.

```r
sum(is.na(train_df))
```

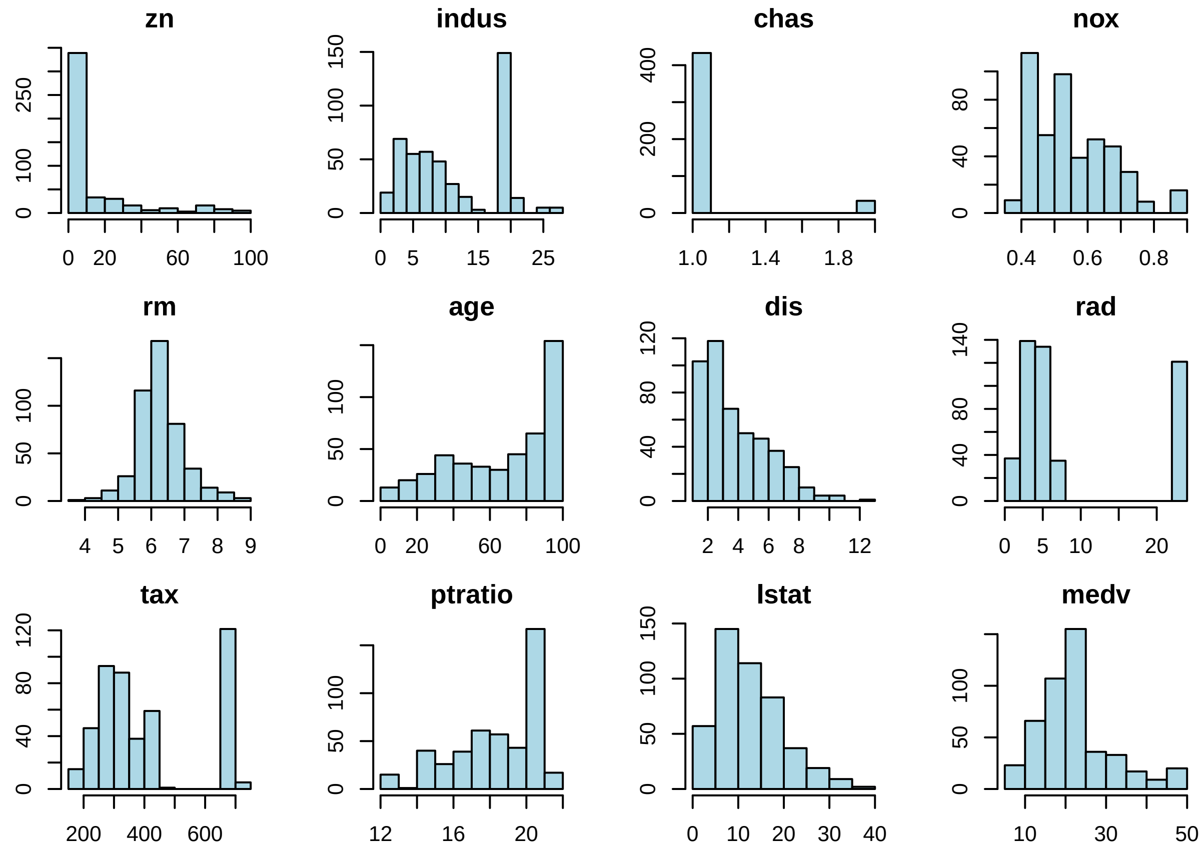
```
## [1] 0
```

There are in fact no missing values in the dataset.

Let's take a look at the distributions for the predictor variables.

```r
par(mfrow=c(3,4))
par(mai=c(.3,.3,.3,.3))

variables <- names(train_df)
factors <- c("chas", "target")
for (i in 1:(length(variables)-1)) {
    if (variables[i] %in% factors){
        hist(as.numeric(train_df[[variables[i]]]), main = variables[i],
             col = "lightblue")
    }else{
        hist(train_df[[variables[i]]], main = variables[i], col = "lightblue")
    }
}
```



The distribution for `rm` appears to be normal, and the distribution for `medv` is nearly normal. The distributions for `zn`, `dis`, `lstat`, and `nox` are right-skewed. The distributions for `age` and `ptratio` are left-skewed.

The distributions for the remaining variables are multimodal, including the distribution for `chas`, which appears degenerate at first glance. It looks like a near-zero variance predictor, which we can confirm using the `nearZeroVar` function from the `caret` package.

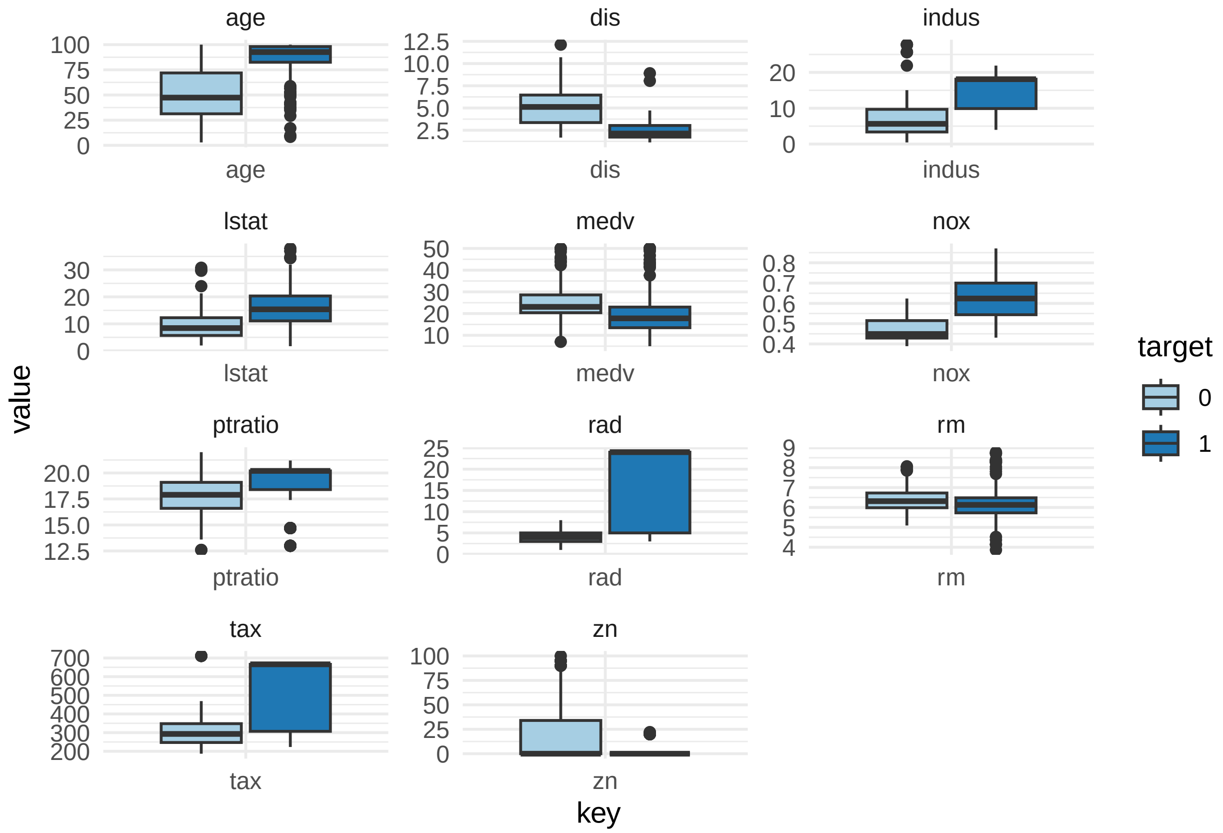
```
nzv <- nearZeroVar(train_df |> select(-target), saveMetrics = TRUE)
knitr::kable(nzv)
```

	freqRatio	percentUnique	zeroVar	nzv
zn	16.142857	5.5793991	FALSE	FALSE
indus	4.321429	15.6652361	FALSE	FALSE
chas	13.121212	0.4291845	FALSE	FALSE
nox	1.176471	16.9527897	FALSE	FALSE
rm	1.000000	89.9141631	FALSE	FALSE
age	10.500000	71.4592275	FALSE	FALSE
dis	1.000000	81.5450644	FALSE	FALSE
rad	1.110092	1.9313305	FALSE	FALSE
tax	3.457143	13.5193133	FALSE	FALSE
ptratio	4.000000	9.8712446	FALSE	FALSE
lstat	1.000000	90.9871245	FALSE	FALSE
medv	2.142857	46.7811159	FALSE	FALSE

The percentage of unique values, `percentUnique`, in the sample for this predictor is less than the typical threshold of 10 percent, but there is a second criterion to consider: the `freqRatio`. This measures the frequency of the most common value (0 in this case) to the frequency of the second most common value (1 in this case). The `freqRatio` value for this predictor is less than the typical threshold of 19 (i.e. 95 occurrences of the most frequent value for every 5 occurrences of the second most frequent value). So it is not considered a near-zero variance predictor. Neither are any of the other predictors.

Next we analyze boxplots to determine the spread of the numeric predictor variables. This will also reveal any outliers.

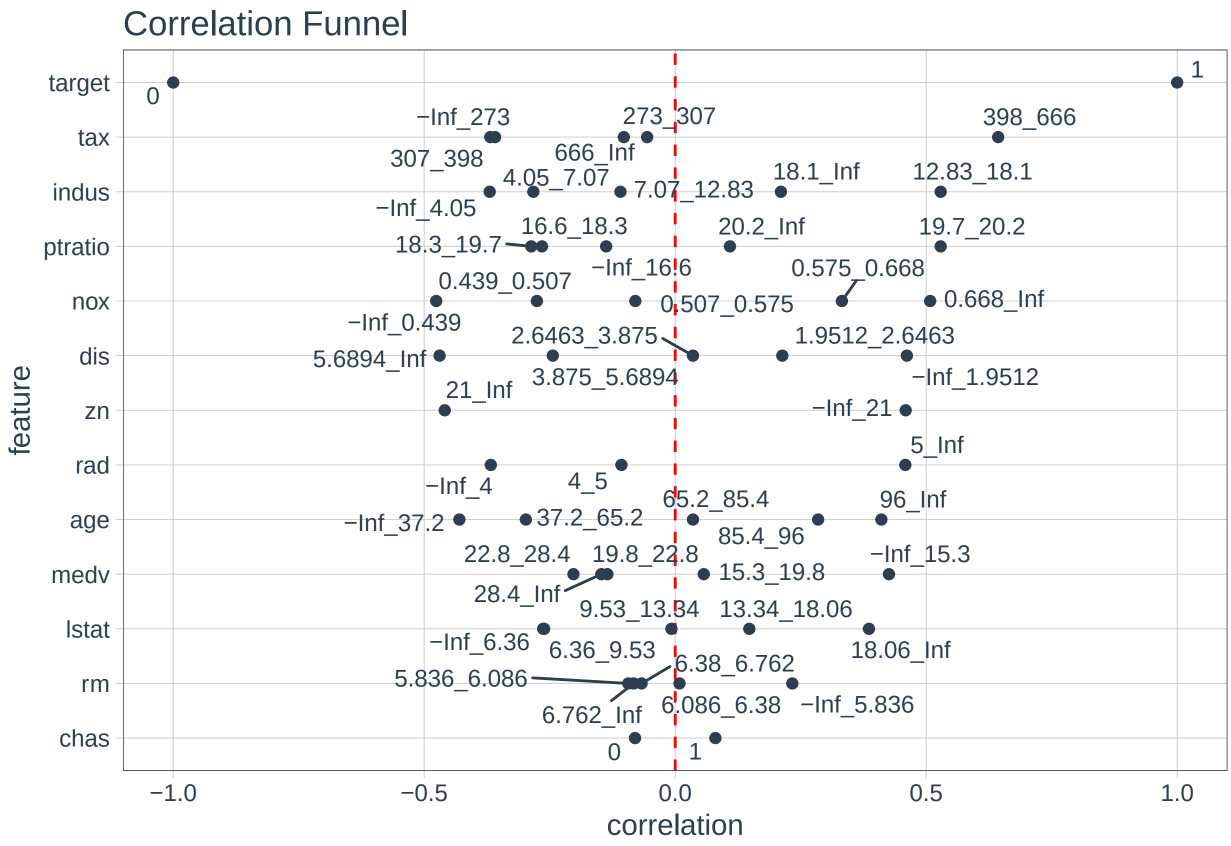
```r
train_df |>
    dplyr::select(-chas) |>
    gather(key, value, -target) |>
    mutate(key = factor(key),
           target = factor(target)) |>
    ggplot(aes(x = key, y = value)) +
    geom_boxplot(aes(fill = target)) +
    facet_wrap(~ key, scales = 'free', ncol = 3) +
    scale_fill_brewer(palette = "Paired") +
    theme_minimal()
```



For certain predictors, the variance between the two categories of the response variable differs largely: `age`, `dis`, `nox`, `rad`, and `tax`.

Next we produce a correlation funnel to visualize the strength of the relationships between our predictors and our response.

```
train_df_binarized <- train_df |>
    binarize(n_bins = 5, thresh_infreq = 0.01, name_infreq = "OTHER",
           one_hot = TRUE)
train_df_corr <- train_df_binarized |>
    correlate(target__1)
train_df_corr |>
    plot_correlation_funnel()
```



The correlation funnel plots the most important features towards the top. In our dataset, the four most important features correlated with the response variable are `tax`, `indus`, `ptratio`, and `nox`.

Looking at the features towards the bottom, the variable `chas` is the least correlated to `target` by the Pearson Correlation coefficient. The correct coefficient to use to understand the strength of the relationship between two categorical variables is actually the ϕ coefficient. If one of the categorical variables had more than two categories, we would need to calculate ϕ using the formula for Cramer's V (also called Cramer's ϕ). However, in the special case that both categorical variables are binary, the value of the Cramer's V coefficient will actually be equal to the value of the Pearson Correlation coefficient. So either formula actually results in the same value for ϕ. We prove this below.

```
cramersv <- round(cramersv(train_df |> select(all_of(factors))), 5)
pearson <- round(cor(as.numeric(train_df$chas), as.numeric(train_df$target), method = "pearson"), 5)
(cramersv == pearson)
```

```
## [1] TRUE
```

The value for ϕ is 0.08004 regardless of the formula used to calculate it, and this very low value indicates very little correlation between `chas` and `target`.