Predict if customer defaults - Default data

Naga Pakalapati

Predict if customer defaults

Default dataset from **ISLR** library contains information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt. It is a four-dimensional dataset with 10000 observations. The question of interest is to predict individuals who will default . We want to examine how each predictor variable is related to the response (default). We shall do the following on this dataset.

- Perform descriptive analysis on the dataset to have an insight. Use summaries and appropriate exploratory graphics to answer the question of interest.
- Use R to build a logistic regression model.
- Discuss result. Which predictor variables were important? Are there interactions?
- How good is the model? We will assess the performance of the logistic regression classifier and find the error rate?

Descriptive Analysis

Dataset Intro: This simulated dataset contains information on ten thousand customers (rows) and the following 4 variables (columns).

- **default**: A factor with levels No and Yes indicating whether the customer defaulted on their debt.
- **student**: A factor with levels No and Yes indicating whether the customer is a student
- **balance**: The average balance that the customer has remaining on their credit card after making their monthly payment
- **income**: Income of customer

Table 1: First few rows of data

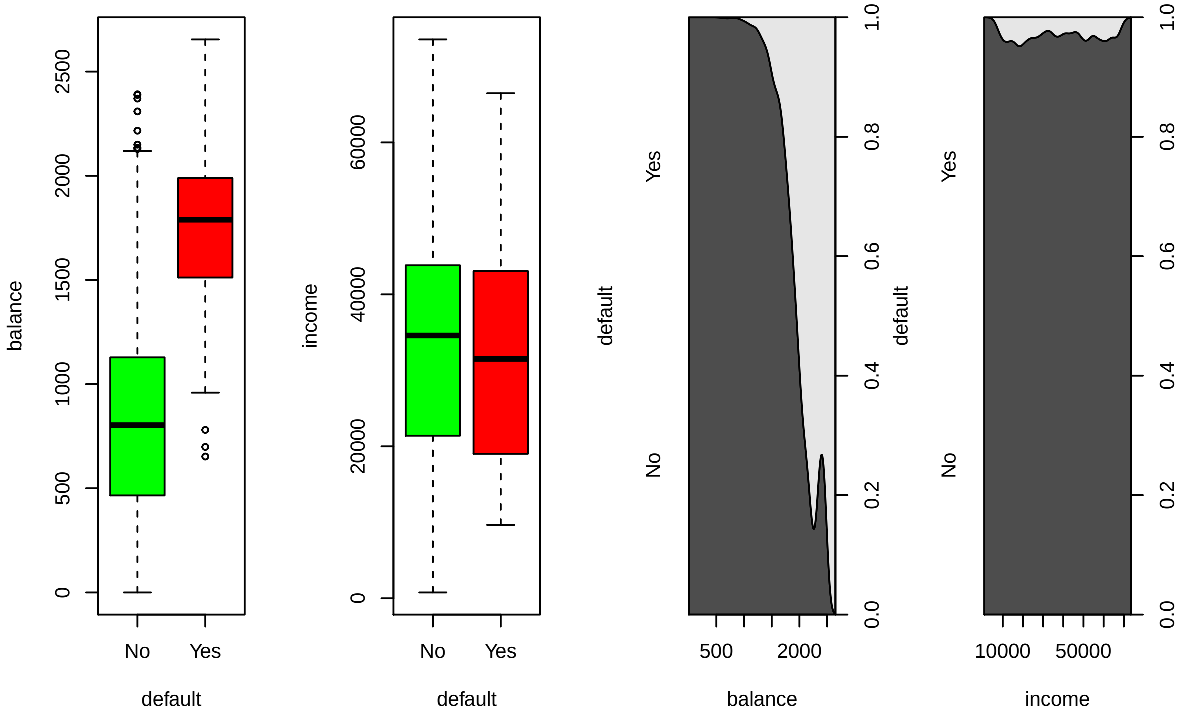
default	student	balance	income
No	No	729.5265	44361.625
No	Yes	817.1804	12106.135
No	No	1073.5492	31767.139
No	No	529.2506	35704.494
No	No	785.6559	38463.496
No	Yes	919.5885	7491.559

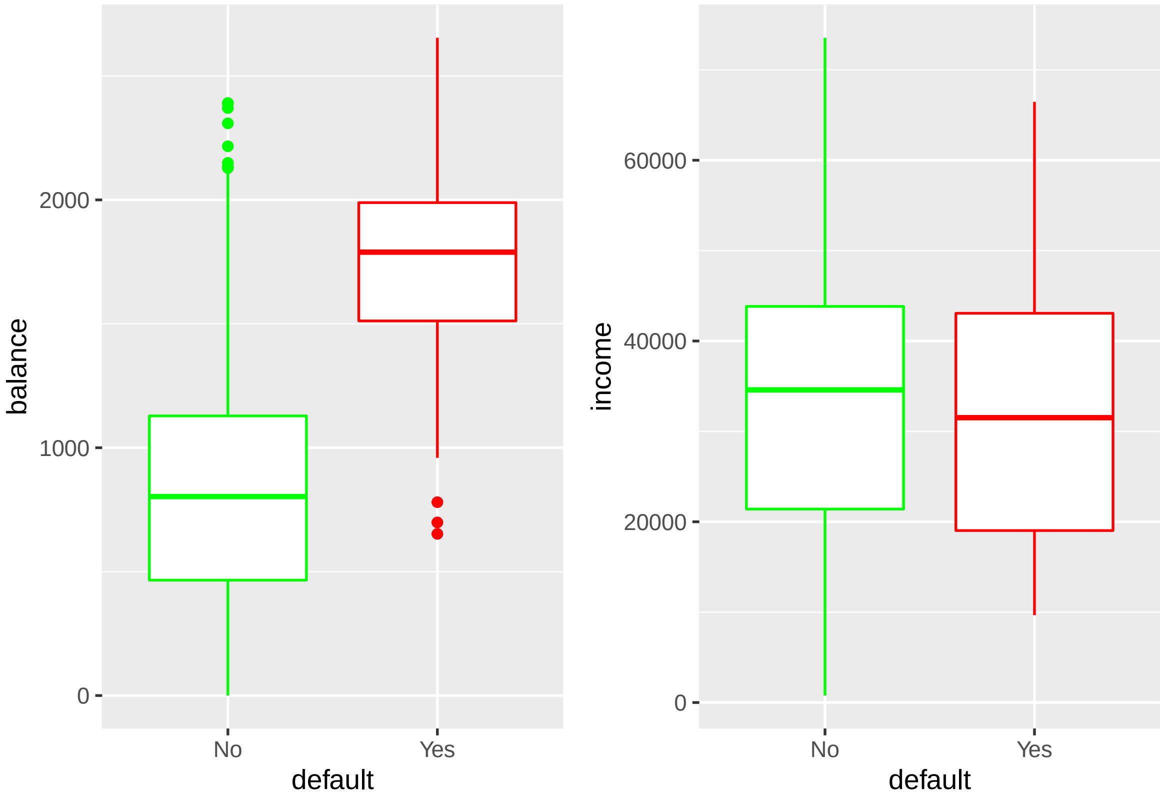
The **default** column is the response variable while the other 3 columns **student**, **balnce** and **income** are the explanatoty variables. Let's check the frequency counts and distributions of the variables. And also check if there are any missing values in the dataset.

```
##  default    student       balance          income     
##  No :9667   No :7056   Min.   :   0.0   Min.   :  772  
##  Yes: 333   Yes:2944   1st Qu.: 481.7   1st Qu.:21340  
##                        Median : 823.6   Median :34553  
##                          Mean   : 835.4   Mean   :33517  
##                          3rd Qu.:1166.3   3rd Qu.:43808  
##                          Max.   :2654.3   Max.   :73554
```

```
## Number of rows with missing values in the data: 0
```

The dataset is good for exploratory analysis. Let's check if we explain **default** column by the variability in balance and income, like if low balance group will default less compared to high or high income will default less compared to low inocome group.





We see a positive evidence that customers who defualt usually have high average monthly balance post payments. In general, this is to be true as making small monthly payments will lead to high balances and interests piling up. The longer this trend continues and the more the balance on the card to pay off, the higher the chances are to default.

However, the income variable doesn't explain dafualt status much. The defaulters and the non defaulters fall in simmilar income range, with non defaulters being on slightly larger range (both low and high end) but it is not very significant.

Now, let's check if being a student has any effect on default.

	Total	Defaulters	Non-Defaulters
No	70.56	61.86186	70.85963
Yes	29.44	38.13814	29.14037

The above table shows the percentages of non students (No) and students (Yes) in entire data set, defaulters and non-defaulters. There is a slight increase in total percentage of students in defaulters category compared to non-defaulters but not so significant.

Let's now build a logistic regression model and evaluate it.

```
##
## Call:
## glm(formula = model4, family = binomial(), data = Default)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.4578  -0.1422  -0.0559  -0.0203   3.7435
##
```

```
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.075e+01  3.692e-01 -29.116  < 2e-16 ***
## studentYes  -7.149e-01  1.475e-01  -4.846 1.26e-06 ***
## balance      5.738e-03  2.318e-04  24.750  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2920.6  on 9999  degrees of freedom
## Residual deviance: 1571.7  on 9997  degrees of freedom
## AIC: 1577.7
##
## Number of Fisher Scoring iterations: 8
```

After iterating over several models, we found that **balance** is highly significant predictor as we discussed earlier and suprisingly being student is less likely to default than non-student. We have removed the income from the model as it's impact on the predicted variable is no significant. Also we don't find any significant interaction variables.

Now let's use this model to train/test the data to predict outcome and evaluate the accuracy of the model.

- We will split the data into train and test sets (75/25).
- Train the model on train and test the results on test datasets
- Campare the predictions with original test data results and
- calculate model accuracy

```
## Proportion of correctly classified observations: 96.96 %
```

```
## Error rate: 3.04 %
```

Our model prediction is very high with a low error rate. This can be considered as a good model. Given the data, 97% of the times we are able to accurately predict if a customer will default or not.