Logistic Regression. Predicting Credit Card Default

Diego Ramos Crespo

May 7, 2025

a) Descriptive analysis

In the current document logistic regression is used to predict whether an individual will default on their credit card payments or not. The prediction is based on three variables: annual income, monthly credit card balance, and whether the individual is a student. The data for this study comes from the `'Default'` dataset in the `'ISLR'` package in `R`. It contains 3 feature columns that provide information about the person. They include two float variables that include the income and balance, plus a categorical variable describing wether the person is a student or not. Finally, the data set contains the target binary column called `'Default'`.

In order to understand the distribution, it is clever to begin with a descriptive analysis which provide some information about the relationships among the variables. Visualizations help assess how the balance and income vary between defaulters and non-defaulters, and whether these features may be useful for predicting default behavior.

Table 1 shows summary statistics of the `Default` dataset. The sample contains 10,000 individuals, with 2944 students and 7056 non-students. However, we can also see a class imbalance between defaulters, with 9667 of people that are marked with this status and 333 non-defaulters. This represents an approximately 1:29 ratio, which can bias our metrics.

Table 1: Summary statistics of the `Default` dataset

Variable	**Summary**
`default`	No: 9667, Yes: 333
`student`	No: 7056, Yes: 2944
`balance`	Min: 0, Median: 823.6, Mean: 835.4, Max: 2654.3
`income`	Min: 772, Median: 34553, Mean: 33517, Max: 73554

Figure 1 display graphical summaries of the variables. The scatter plot shows a clear trend, where individuals with higher credit card balances are more likely to default. In contrast, annual income appears to be less predictive of default status.

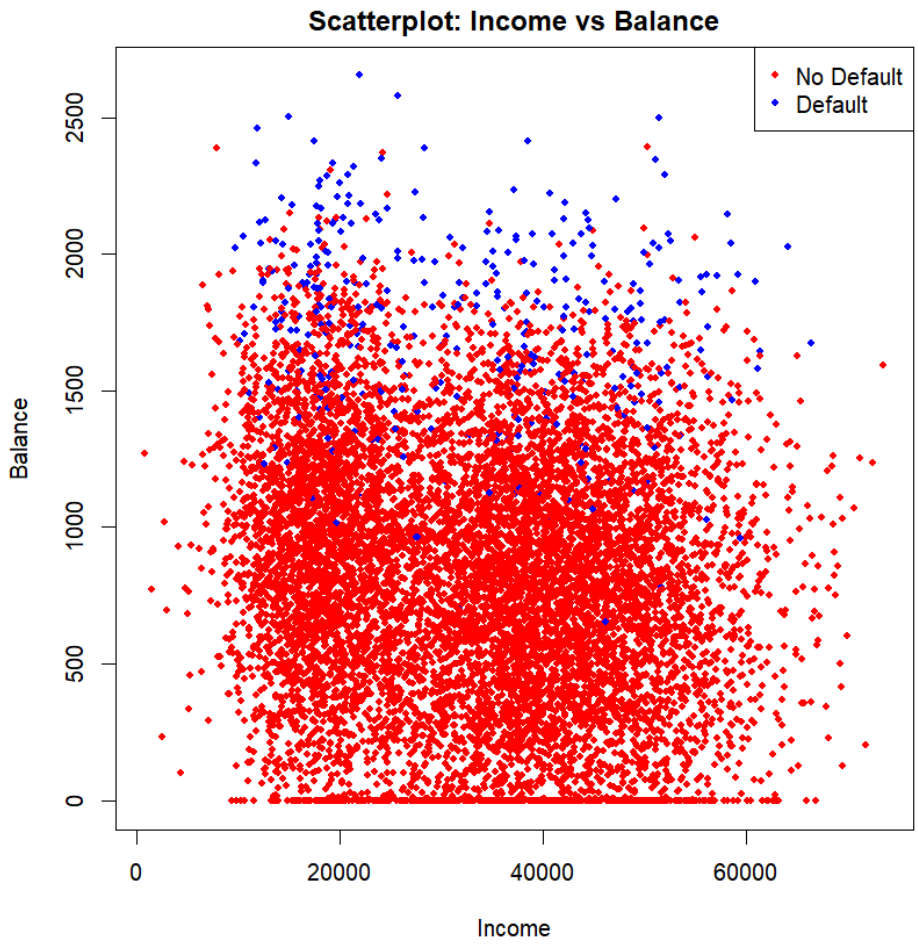

Figure 1: Income vs. Balance colored by default status

Figure 2 presents a comparative view of credit card holders based on their default status. The left panel highlights individuals who did not default, while the right panel shows those who did. A clear trend is observed: individuals with higher balance amounts are more likely to default on their credit card payments, regardless of income level. In contrast, those with lower balances are less likely to default. This suggests that the credit card balance may be a stronger predictor of default than income. Although income varies widely in both groups, balance levels show a more distinct pattern associated with default behavior.

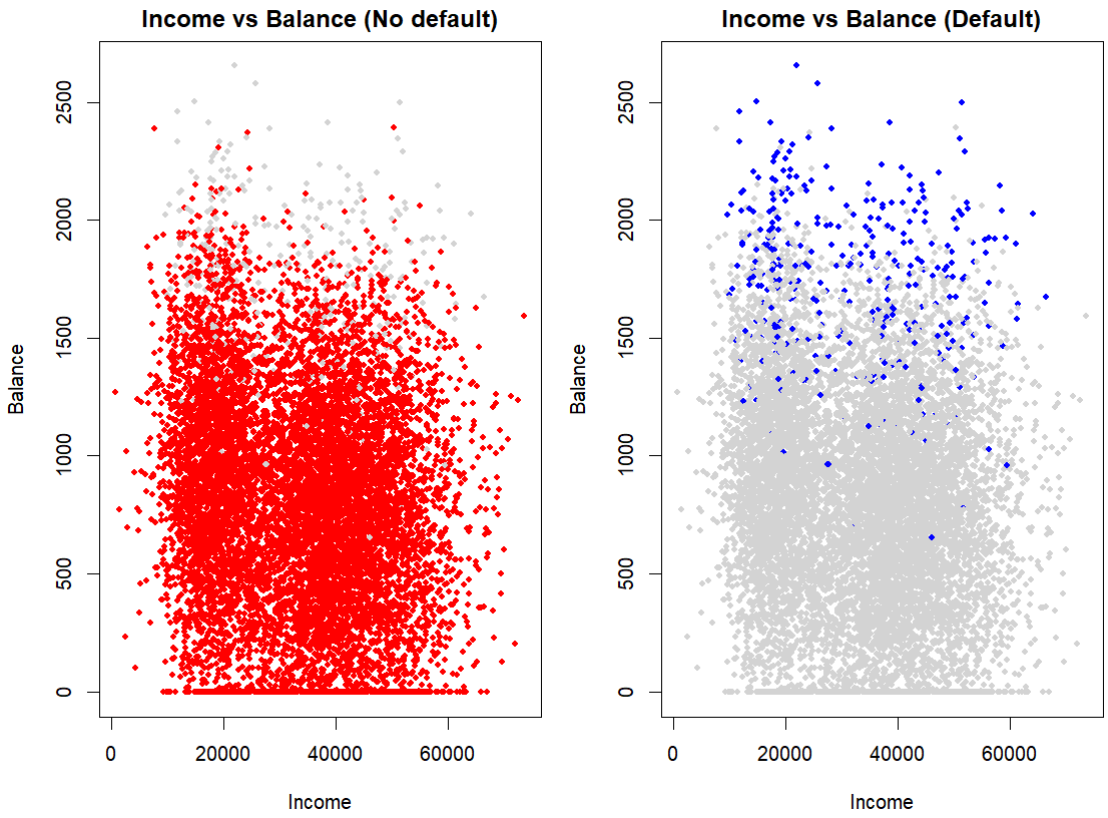

Figure 2: Boxplots of Balance and Income grouped by Default status

Figure 3 shows boxplots comparing the distributions of `Balance` and `Income` between individuals who defaulted and those who did not. The left panel clearly demonstrates that defaulters tend to have significantly higher credit card balances, with a higher median and a more spread-out distribution. In contrast, the right panel indicates that income distributions are relatively similar across both default groups, with only slight differences in central tendency and variability. These visual patterns reinforce the idea that balance is a more influential factor than income in predicting the likelihood of default.

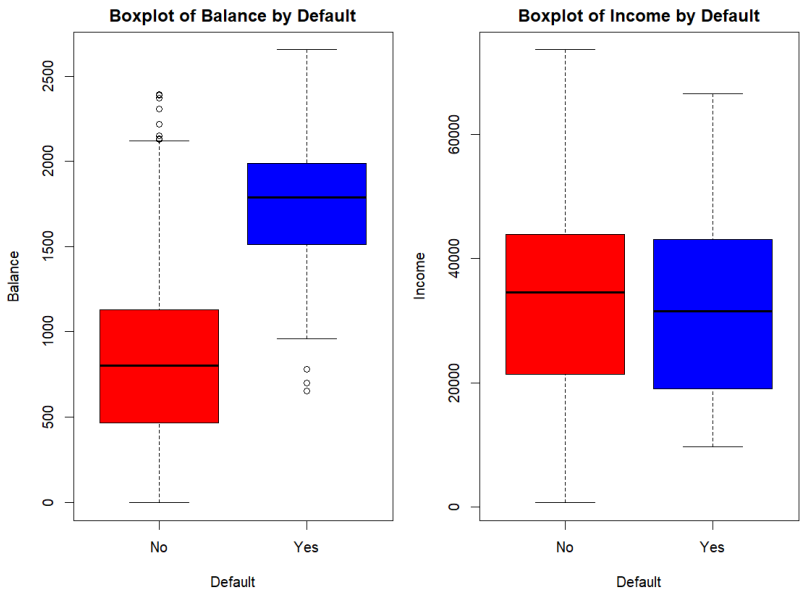

Figure 3: Boxplots of Balance (left) and Income (right) grouped by default status.

Figure 4 displays a conditional density plot of `default` given `balance`. The black region corresponds to the conditional probability of `default = "Yes"`, whereas the gray region represents `default = "No"`.

We can observe that for lower balance values, the likelihood of default is minimal, as shown by the predominance of the gray area. However, as the balance increases, the probability of default also increases, indicated by the expanding black region. At higher balances (above $2,000), the likelihood of default becomes higher, suggesting that customers with larger balances are more prone to be in a default status. This plot effectively visualizes how the conditional probability of default varies as a function of balance.

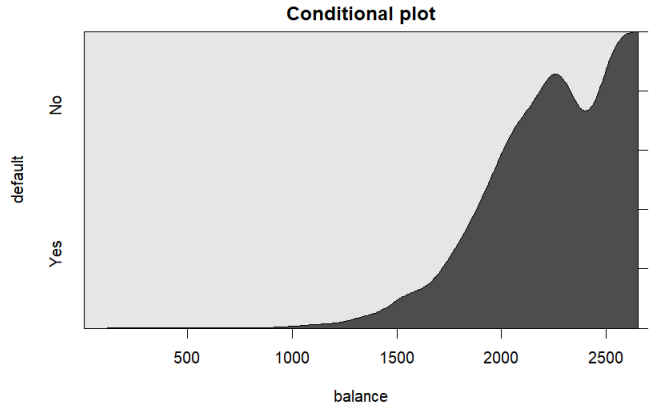

Figure 4: Boxplots of Balance (left) and Income (right) grouped by default status.

b) Predicting default based on student status

We fit a logistic regression model to predict the probability of credit card default based on the customer's balance. The values are displayed in 2

Table 2: Logistic Regression Output: `default` $\sim$ `balance`

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(¿|z|)**
Intercept	-10.6513	0.3612	-29.49	$< 2 \times 10^{-16}$
balance	0.005499	0.0002204	24.95	$< 2 \times 10^{-16}$

The estimated coefficient for `balance` is $\hat{\beta}_1 = 0.005499$, which represents the change in the log-odds of default for a one-unit increase in balance.

This means that for a one-dollar increase in balance, the odds of default increase by a factor of:

$$e^{0.005499} \approx 1.0055$$

This implies that each additional dollar in balance increases the odds of default by approximately 0.55%.

The probability of default significantly depends on `balance`. This is supported by the very small p-value ($< 2 \times 10^{-16}$) associated with the `balance` coefficient in the model output. Since this p-value is much smaller than any reasonable significance level (e.g., $\alpha = 0.05$), we reject the null hypothesis that the coefficient is zero. This provides strong evidence that `balance` is a statistically significant predictor of `default`.

Next, the predicted probabilities of default for balances of \$950, \$1550, and \$1990 are computed using the fitted model:

- Balance = \$950: $\hat{P}(\text{default}) = 0.0035$
- Balance = \$1550: $\hat{P}(\text{default}) = 0.1160$
- Balance = \$1990: $\hat{P}(\text{default}) = 0.4473$

As expected, the probability of default increases sharply with higher balances. A balance of \$950 corresponds to a very low default probability, while a balance of \$1990 corresponds to a nearly 45% chance of default, highlighting the predictive strength of the balance variable.

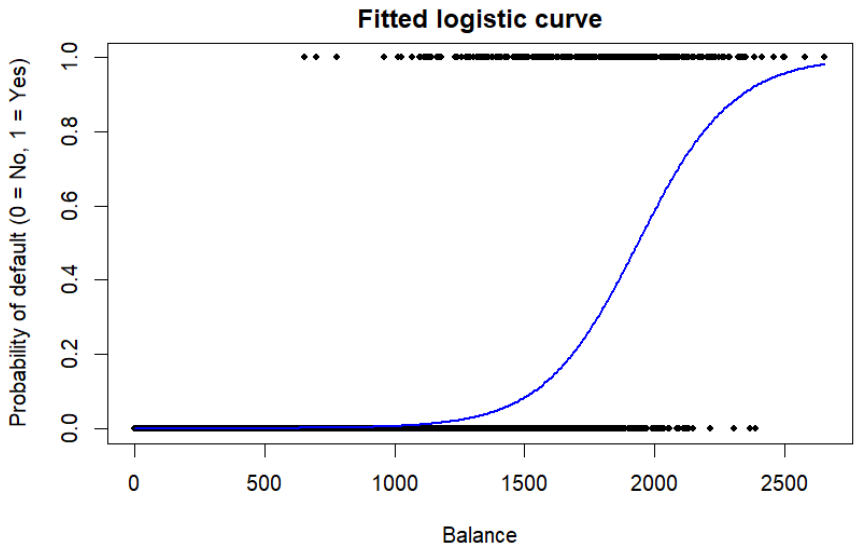

Figure 5: Fitted logistic regression curve for default vs. balance.

Figure 5 shows the actual data points for default (coded as 0 = No, 1 = Yes) and the fitted logistic regression curve. The increasing trend clearly indicates that the likelihood of default grows as the balance increases.

c) Predicting default based on student status

We now examine whether being a student influences the probability of default by fitting a logistic regression model with `student` as the only predictor. The logistic regression model output is obtained as shown in Table 3.

Table 3: Logistic Regression Output: `default` $\sim$ `student`

Coefficient	**Estimate**	**Std. Error**	**p-value**
Intercept	-3.50413	0.07071	$< 2 \times 10^{-16}$
studentYes	0.40489	0.11502	0.000431

From Table 3, we observe that the coefficient for `studentYes` is statistically significant ($p = 0.000431 < 0.05$). This suggests that being a student has a statistically significant effect on the probability of default.

The estimated logistic regression equation is:

$$\log\left(\frac{P(\text{default})}{1-P(\text{default})}\right) = -3.50413 + 0.40489 \cdot \text{student}$$

where `student` is coded as 1 for students and 0 for non-students.

c.1) Predicted Probabilities

To better understand the impact, we calculate the predicted probabilities of default for students and non-students:

- **Non-student (student = 0):**

$$P(\text{default}) = \frac{1}{1+\exp(-(-3.50413))} \approx 0.0292$$

- **Student (student = 1):**

$$P(\text{default}) = \frac{1}{1+\exp(-(-3.50413+0.40489))} \approx 0.0438$$

Although students show a higher estimated probability of default than non-students (4.38% vs 2.92%), the overall probabilities remain low. However, the effect is statistically significant, indicating that student status does influence the likelihood of default.

Thus, we conclude that being a student slightly increases the probability of default, and this effect is statistically significant based on the regression model.

d) Logistic Regression model with balance, student and income

To better predict the probability of default, we now include three predictors in the logistic regression model: balance, student, and income (in thousands).

The summary output of the model is presented in the Table 4.

Table 4: Logistic Regression Output: `default ~ balance + student + income`

Variable	**Estimate**	**Std. Error**	**z value**	**Pr(¿|z|)**
Intercept	-10.87	0.4923	-22.08	$< 2 \times 10^{-16}$
balance	0.005737	0.0002319	24.738	$< 2 \times 10^{-16}$
studentYes	-0.6468	0.2363	-2.738	0.00619
income	3.033×10^{-6}	8.203×10^{-6}	0.370	0.71152

Now, based on the p-values displayed in the 4 provide significant information.

- The coefficient for **balance** is highly significant ($p < 0.001$), suggesting that higher balances strongly increase the probability of default.
- The coefficient for **studentYes** is significant at the 1% level ($p = 0.00619$), indicating that being a student has a statistically significant effect on the probability of default, although the coefficient is negative, implying students are less likely to default, mantaining other variables constant.
- The coefficient for **income** is *not significant* ($p = 0.71152$), meaning that income (in thousands) does not have a statistically meaningful impact on the probability of default after controlling for the other variables.

In this model, balance and student are significant predictors of the log-odds of default, while income is not. Therefore, we can say that only two of the three variables are useful for explaining the likelihood of default in this context.

e) Predicting default based on balance and student status

To analyze the probability of default based on credit card balance and student status, we fit a logistic regression model using these two predictors. The summary of the model output is shown in Table 5.

Table 5: Logistic Regression Output: `default` ∼ `balance + student`

Variable	Estimate	Std. Error	z value	Pr(¿|z|)
Intercept	−10.75	0.3692	−29.116	$< 2 \times 10^{-16}$
balance	0.005738	0.0002318	24.750	$< 2 \times 10^{-16}$
studentYes	−0.7149	0.1475	−4.846	1.26×10^{-6}

From Table 5, we observe that both `balance` and `student` are statistically significant predictors of default. The coefficient for `balance` is positive and highly significant ($p < 0.001$), indicating that as balance increases, the probability of default increases. The coefficient for `studentYes` is negative and also highly significant ($p = 1.26 \times 10^{-6}$), suggesting that students are less likely to default than non-students, holding balance constant.

The estimated logistic regression equation is the one presented below.

$$\log\left(\frac{P(\text{default})}{1 - P(\text{default})}\right) = -10.75 + 0.005738 \cdot \text{balance} - 0.7149 \cdot \text{student}$$

where `student` is 1 for students and 0 for non-students.

e.1) Probability of default by student status

To visualize the model, we plot the predicted probability of default as a function of `balance`, separately for students and non-students. The Figure 6 below shows the resulting curves.

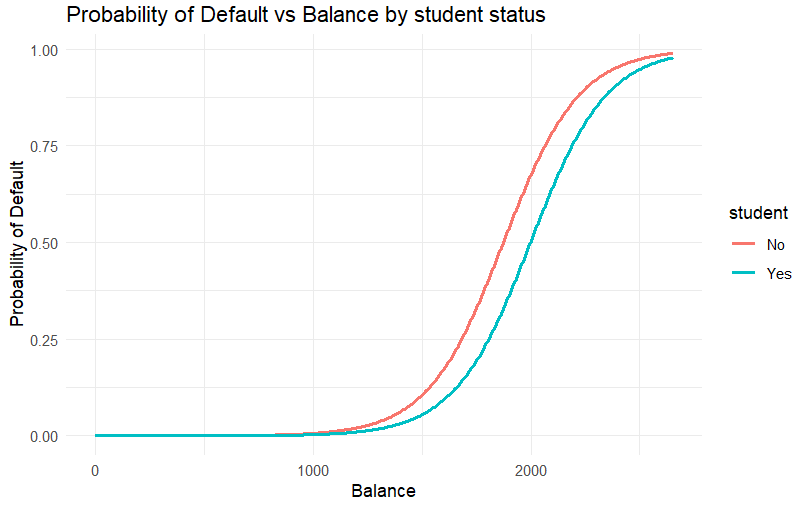

Figure 6: Probability of default by balance for students and non-students

Figure 6 illustrates the relationship between `balance` and the predicted probability of default for both student categories. We can clearly observe that:

- For any given level of balance, students tend to have a lower probability of default compared to non-students.
- As balance increases, the probability of default rises steeply for both groups, but the curve for non-students consistently lies above the student curve.
- The difference in predicted probabilities between students and non-students becomes more pronounced at higher balances.

These findings confirm that although higher balances are strongly associated with an increased risk of default, student status appears to mitigate this risk. Therefore, non-students are more likely to default than students with the same balance level.

f) Conclusion: Single vs Two-predictor logistic models

In comparing the logistic regression models with one predictor (student) and two predictors (balance and student), several important conlusions can be drawn:

- **Single Predictor Model (Student Only):** As shown in Table 3, the variable `student` was found to be statistically significant ($p = 0.000431$), with students exhibiting a slightly higher probability of default than non-students. The predicted probability of default was approximately 2.92% for non-students and 4.38% for students. Although statistically significant, the overall predictive power of the model is limited since it relies on a single categorical variable with a small effect size.
- **Two-Predictor Model (Balance + Student):** In contrast, the model shown in Table 5 includes both `balance` and `student` as predictors. Here, both variables are statistically significant, but the interpretation of `student` changes:
  - Balance remains a strong, positive, and highly significant predictor of default.
  - Student = Yes becomes negative and remains statistically significant ($p = 1.26 \times 10^{-6}$), implying that—after controlling for balance—students are actually *less* likely to default compared to non-students.

The change in the direction of the `student` coefficient from positive (in the single-predictor model) to negative (in the two-predictor model) suggests the presence of a confounding relationship between student and `balance`. That is, students tend to carry lower balances on average, which is associated with a lower default risk. When balance is not accounted for, students appear more likely to default. Once we control for balance, we see that—holding balance constant—students are actually at lower risk.

- **Model Comparison:** The model using two predictors provides a more accurate and nuanced view of the factors influencing default. It adjusts for the imbalance in `balance` distributions across student status, allowing for a better understanding of the independent effect of being a student.

While the `student` variable is statistically significant in both models, its effect is better understood in the context of the two-predictor model. Including `balance` provides essential context that reveals students are, in fact, less likely to default once balance is taken into account. Therefore, the model with two predictors is superior for understanding and predicting credit card default behavior.

g) Evaluating model Performance with two Predictors

In this section, the performance of the model is evaluated using a a logistic regression model using `balance` and `student` as predictors. A random sample of 5000 observations from the `Default` dataset was selected as the training set, and the remaining observations formed the test set.

g.1 Accuracy with threshold = 0.5

The model was trained on the training set and used to predict probabilities of default on the test set. Using a classification threshold of $p = 0.5$, we classify an observation as "Yes" (Default) if $p > 0.5$ and "No" otherwise.

We compare these predictions with the actual default values in the test set. The proportion of correctly classified observations (accuracy) is:

$$\text{Accuracy} = \frac{\text{Number of correct predictions}}{\text{Total number of test observations}} = 0.954 \quad (95.4\%)$$

This indicates that the model correctly classifies 95.4% of the test cases using only the `balance` and `student` variables.

g.2 Effect of changing the classification threshold

To evaluate whether the performance improves with a different threshold, we tested two alternative values:

- Threshold = 0.3:

$$\text{Accuracy} = 0.9402 \quad (94.02\%)$$

- Threshold = 0.7:

$$\text{Accuracy} = 0.9608 \quad (96.08\%)$$

Changing the threshold affects the classification performance. A lower threshold increases the number of predicted defaults but may lead to more false positives, reducing overall accuracy. In contrast, a higher threshold leads to fewer defaults being predicted, but in this case, it slightly improves accuracy.

Although changing the threshold from 0.5 does lead to different accuracy values, accuracy alone may not be the best metric, especially if the dataset is imbalanced. Further evaluation using metrics such as precision, recall, or the ROC curve may be more useful.