sources of bias, especially when the dataset only contains a subset of the observations of interest.

The *Parole.csv* dataset contains all individuals released from parole in 2004, either due to completing their parole term or violating the terms of their parole. However, it does not contain parolees who neither violated their parole nor completed their term in 2004, causing non-violators to be underrepresented. This is called *selection bias* or *selecting on the dependent variable*, because we used our dependent variable (parole violation) to select only a subset of all relevant parolees to include in our analysis. How could we improve our dataset to best address selection bias?

Loan Repayment

In the lending industry, investors provide loans to borrowers in exchange for the promise of repayment with interest. If the borrower repays the loan, then the lender profits from the interest. However, if the borrower is unable to repay the loan, then the lender loses money. Therefore, lenders would like to minimize the risk of a borrower being unable to repay a loan.

In this exercise, we will use publicly available data from LendingClub, a website that connects borrowers and investors over the internet. The dataset is in the file *Loans.csv* (available in the Online Companion). There are 9,578 observations, each representing a 3-year loan that was funded through the LendingClub.com platform between May 2007 and February 2010. There are 14 variables in the dataset, described in Table 22.6. We will be trying to predict `NotFullyPaid`, using all of the other variables as independent variables.

a) Let us start by building a logistic regression model.

i) First, randomly split the dataset *Loans.csv* into a training set and a testing set. Put 70% of the data in the training set. What is the accuracy on the test set of a simple baseline model that predicts that all loans will be paid back in full (`NotFullyPaid` = 0) ? Our goal will be to build a model that adds value over this simple baseline method.

ii) Now, build a logistic regression model that predicts the dependent variable `NotFullyPaid` using all of the other variables as independent variables. Use the training set as the data for the model. Describe your resulting model. Which of the independent variables are significant in your model?

iii) Consider two loan applications, which are identical other than the fact that the borrower in Application A has a FICO credit score of 700 while the borrower in Application B has a FICO credit score of 710. Let Logit(A) be the value of the linear logit

Table 22.6: Variables in the dataset *Loans.csv*.

Variable	Description
`CreditPolicy`	1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise.
`Purpose`	The purpose of the loan (one of "Credit Card," "Debt Consolidation," "Educational," "Major Purchase," "Small Business," or "Other").
`IntRate`	The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Note that borrowers judged by LendingClub to be more risky are assigned higher interest rates.
`Installment`	The monthly installments ($) owed by the borrower if the loan is funded.
`LogAnnualInc`	The natural log of the self-reported annual income of the borrower.
`Dti`	The debt-to-income ratio of the borrower (amount of debt divided by annual income).
`Fico`	The FICO credit score of the borrower.
`DaysWithCrLine`	The number of days the borrower has had a credit line.
`RevolBal`	The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).
`RevolUtil`	The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).
`InqLast6mths`	The borrower's number of inquiries by creditors in the last 6 months.
`Delinq2yrs`	The number of times the borrower had been 30+ days past due on a payment in the past 2 years.
`PubRec`	The borrower's number of derogatory public records (bankruptcy fillings, tax liens, or judgments).
`NotFullyPaid`	1 if the loan was not paid back in full, and 0 otherwise.

function of loan A not being paid back in full, according to our logistic regression model, and define Logit(B) similarly for loan B. What is the value of Logit(A) - Logit(B)?

iv) Now predict the probability of the test set loans not being paid back in full. Store these predicted probabilities in a variable named **`PredictedRisk`** and add it to your test set (we will use this variable in later parts of the problem). What is the accuracy of the logistic regression model on the test set using a threshold of 0.5? How does this compare to the baseline model?

v) What is the test set AUC of the model? Given the accuracy and the AUC of the model on the test set, do you think this model could be useful to an investor to make profitable investments?

b) LendingClub assigns the interest rate to a loan based on their estimate of that loan's risk. This variable, **`IntRate`**, is an independent variable in our dataset. In this part, we will investigate just using the loan's interest rate as a "smart baseline" to order the loans according to risk.

i) Using the training set, build a logistic regression model that predicts the dependent variable **`NotFullyPaid`** using **`IntRate`** as the only independent variable. Is **`IntRate`** significant in this model? Was it significant in the first logistic regression model you built? How would you explain this difference?

ii) Use the model you just built (with only one independent variable) to make predictions for the observations in the test set. What is the highest predicted probability of a loan not being paid back in full on the test set? How many loans would we predict would not be paid back in full if we used a threshold of 0.5 to make predictions?

iii) Compute the test set AUC of the model. How does this compare to the model using all of the independent variables? In your opinion, which model is stronger? Why?

c) Let us now see how our logistic regression model can be used to identify loans that are expected to be profitable.

i) If the loan is paid back in full, then the investor makes interest on the loan. However, if the loan is not paid back, the investor loses the money invested. Therefore, the investor should seek loans that best balance this risk and reward.

To compute interest revenue, consider a $\$c$ investment in a loan that has an annual interest rate r over a period of t years. Using continuous compounding of interest, this investment pays back $c \times e^{rt}$ dollars by the end of the t years. How much does a \$10

investment with an annual interest rate of 6% pay back after 3 years, using continuous compounding of interest? (HINT: Remember to convert the percentage to a proportion before doing the math.)

ii) While the investment has this value after collecting interest, the investor had to pay c dollars for the investment. What is the profit to the investor if the investment is paid back in full? What is the profit to the investor if the investment is *not* paid back in full?

iii) Compute the profit of a \$1 investment in each loan, and save your result to a variable named **`Profit`**. Keep in mind that the profit computation should depend on the value of the variable **`NotFullyPaid`**. What is the maximum profit of a \$1 investment in any loan in the testing set?

iv) A simple investment strategy of equally investing in all the loans would yield a profit \$20.94 for a \$100 investment. But this simple investment strategy does not leverage the prediction model we built earlier in this problem. Instead, let us analyze an investment strategy in which the investor only purchases loans with a high interest rate (a rate of at least 15%) to maximize return, but amongst these loans selects the ones with the lowest predicted risk of not being paid back in full. We will model an investor who invests \$1 in each of the most promising 100 loans.

First, create a new dataset called **`HighInterest`** consisting of the test set loans with an interest rate of at least 15%. What is the average profit of a \$1 investment in one of these high-interest loans? What proportion of the high-interest loans were not paid back in full?

v) Next, sort the loans in the **`HighInterest`** dataset by the variable **`PredictedRisk`** that we computed earlier in the problem. Create a new dataset called **`SelectedLoans`** that consists of the 100 loans with the smallest values of **`PredictedRisk`**.

What is the profit of an investor who invested \$1 in each of these 100 loans? How many of the 100 selected loans were not paid back in full? How does this compare to the simple strategy of investing in all loans, which yielded a profit of \$20.94 for a \$100 investment?

d) One of the most important assumptions of predictive modeling often does not hold in financial situations, causing predictive models to fail. What do you think this is? As an analyst, what could you do to improve the situation?