Predicting credit card defaults with logistic regression

Introduction

With credit card balances reaching an average of $16,000 per household and student loans becoming the largest source of household debt, it's clear that Americans are financing many aspects of their lives. For credit card lenders interesting in mitigating loses from risky borrowers, the desire to build models that identify high-risk borrowers has never been stronger.

In this lesson and the next, you'll construct predictive models to identify credit customers who are likely to default on their next payment. In this lesson, you'll create logistic regression models, and next time decision trees.

Data

The data come from Taiwanese credit card customers. Use the str (structure) and summary functions to familiarize yourself with the data. There are 23 features and one outcome variable default_next_month, which is a binary variable indicating whether the customer will default in the following month. See the feature definitions in the table below and read more about the Default of Credit Card Clients Data Set.

```
str(default)
```

```
## 'data.frame':    30000 obs. of  24 variables:
##  $ default_next_month: int  1 1 0 0 0 0 0 0 0 0 ...
##  $ limit_bal         : int  20000 120000 90000 50000 50000 50000 500000 100000 140000 20000 ...
##  $ sex               : int  2 2 2 2 1 1 1 2 2 1 ...
##  $ education         : int  2 2 2 2 2 1 1 2 3 3 ...
##  $ marriage          : int  1 2 2 1 1 2 2 2 1 2 ...
##  $ age               : int  24 26 34 37 57 37 29 23 28 35 ...
##  $ pay_sept          : int  2 -1 0 0 -1 0 0 0 0 -2 ...
##  $ pay_aug           : int  2 2 0 0 0 0 0 -1 0 -2 ...
##  $ pay_july          : int  -1 0 0 0 -1 0 0 -1 2 -2 ...
##  $ pay_june          : int  -1 0 0 0 0 0 0 0 0 -2 ...
##  $ pay_may           : int  -2 0 0 0 0 0 0 0 0 -1 ...
##  $ pay_april         : int  -2 2 0 0 0 0 0 -1 0 -1 ...
##  $ bill_sept         : int  3913 2682 29239 46990 8617 64400 367965 11876 11285 0 ...
##  $ bill_aug          : int  3102 1725 14027 48233 5670 57069 412023 380 14096 0 ...
##  $ bill_july         : int  689 2682 13559 49291 35835 57608 445007 601 12108 0 ...
##  $ bill_june         : int  0 3272 14331 28314 20940 19394 542653 221 12211 0 ...
##  $ bill_may          : int  0 3455 14948 28959 19146 19619 483003 -159 11793 13007 ...
##  $ bill_april        : int  0 3261 15549 29547 19131 20024 473944 567 3719 13912 ...
##  $ pay_amt_sept      : int  0 0 1518 2000 2000 2500 55000 380 3329 0 ...
##  $ pay_amt_aug       : int  689 1000 1500 2019 36681 1815 40000 601 0 0 ...
##  $ pay_amt_july      : int  0 1000 1000 1200 10000 657 38000 0 432 0 ...
##  $ pay_amt_june      : int  0 1000 1000 1100 9000 1000 20239 581 1000 13007 ...
##  $ pay_amt_may       : int  0 0 1000 1069 689 1000 13750 1687 1000 1122 ...
##  $ pay_amt_april     : int  0 2000 5000 1000 679 800 13770 1542 1000 0 ...
```

Predictor	Description
default_payment_next_month	Will the customer default next month (Yes = 1, No = 0)?
limit_bal	Household credit limit
sex	1 = male; 2 = female
education	1 = graduate school; 2 = university; 3 = high school; 4 = others

Predictor	Description
marriage	1 = married; 2 = single; 3 = divorced, 0 = others
age	years
pay_april - pay_sept	Payment status from the previous 6 months (April - Sept. 2006). -2 = No consumption; -1 = Paid in full; 0 = The use of revolving credit; 1 = payment delay for one month; 2 = payment delay for two months; . . . ; 8 = payment delay for eight months; 9 = payment delay for nine months and above.
bill_april - bill_sept	Bill statement amount from the previous 6 months (April - Sept. 2006).
pay_amt_april - pay_amt_sept	Previous payment amount from the previous 6 months (April - Sept. 2006)

Instructions

Follow the steps below to familiarize yourself with the credit card default data set and prepare it for a machine learning analysis. You'll start by building a classifier using logistic regression and continue the model building in the next lesson.

1. Read in the data using the `read.delim` function. Then use `ggplot2`, `ggpairs`, and `dplyr` to identify interesting relationships in the data. Write a short description of one interesting pattern you identified.
2. Construct at least one new feature to include in model development. You might choose to create a new feature based on your findings from the exploratory data analysis. Plot the new variable and interpret the result. Use color to a facet to show the relationship between your new feature and the outcome variable `default_next_month`.
3. Use the `createDataPartition` function from the `caret` package to split the data into a training and testing set. Pre-process the data with `preProcess` as needed.
4. Fit at least 3 logistic regression models.
5. Use the `dotplot` function to compare the accuracy of the models you constructed in 4. Which model performed the best in terms of predictive accuracy?