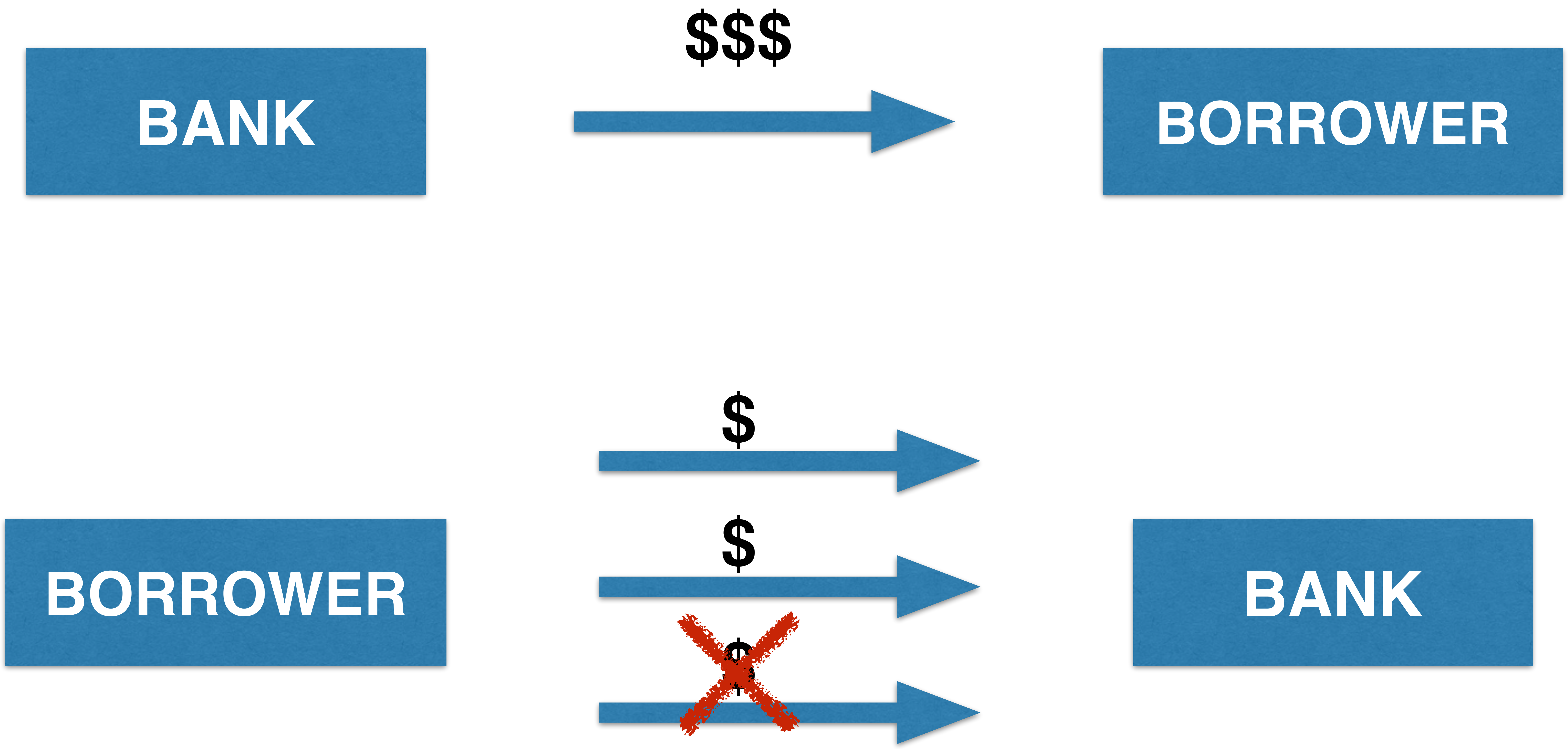
CREDIT RISK MODELING IN R

Introduction and data structure

What is loan default?



Components of expected loss (EL)

- Probability Of default or PD) (%)
- Exposure At default or EAD ($ value)
- Loss given default or LGD (%)

EL= PD x EAD x LGD

Information used by banks

- Application information:
  - income
  - marital status
  - ...
- Behavioral information
  - current account balance
  - payment arrears in account history
  - ...

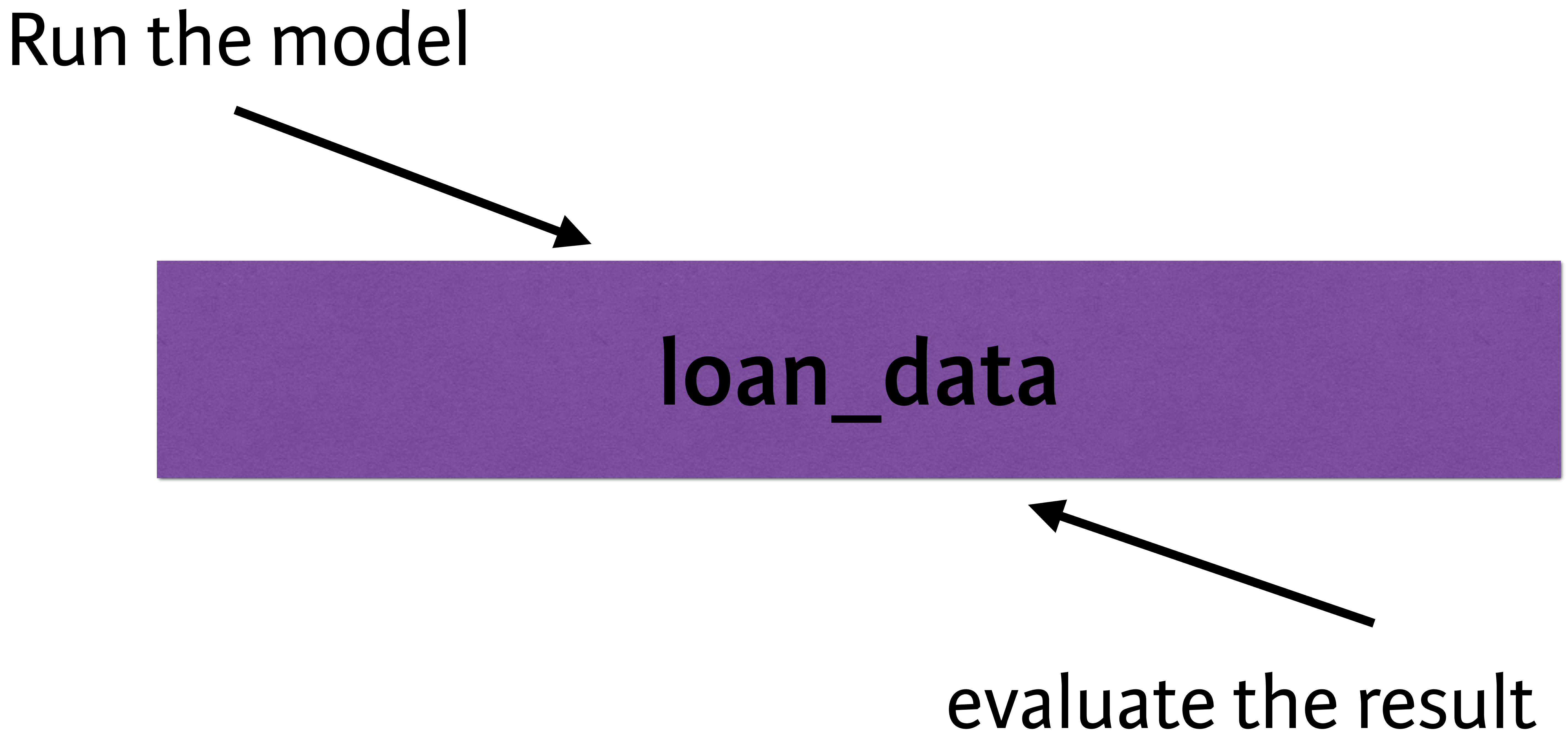
Raw data

```
> head(loan_data, 10)
   loan_status loan_amnt int_rate grade emp_length home_ownership annual_inc age
1            0      5000    10.65     B         10           RENT      24000  33
2            0      2400       NA     C         25           RENT      12252  31
3            0     10000    13.49     C         13           RENT      49200  24
4            0      5000       NA     A          3           RENT      36000  39
5            0      3000       NA     E          9           RENT      48000  24
6            0     12000    12.69     B         11            OWN      75000  28
7            1      9000    13.49     C          0           RENT      30000  22
8            0      3000     9.91     B          3           RENT      15000  22
9            1     10000    10.65     B          3           RENT     100000  28
10           0      1000    16.29     D          0           RENT      28000  22
```

Exploring the data

- Make crosstables, histograms
- Delete/manage outliers
- Manage missing data
    - Delete row/column
    - Replace
    - Keep —> coarse classification (or "binning")

Start analysis

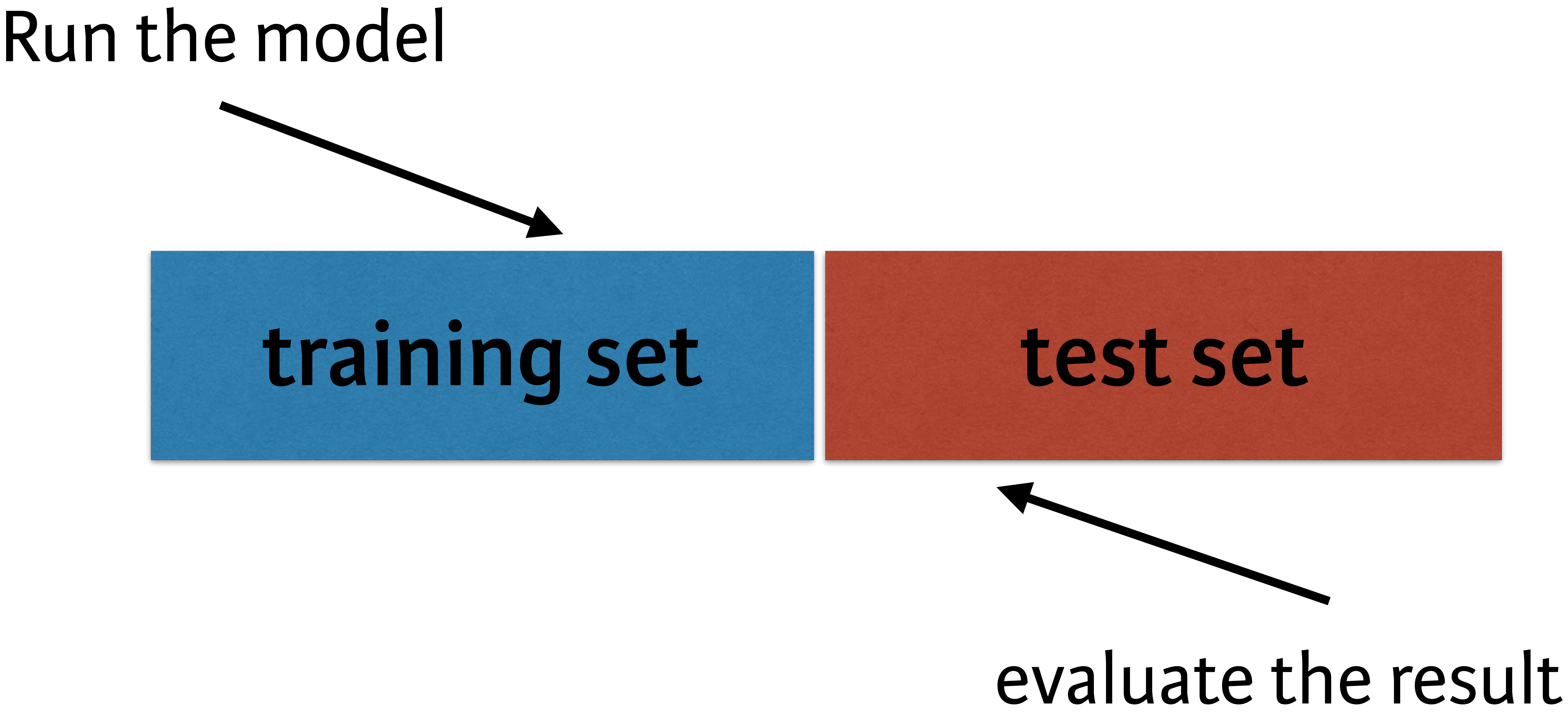


Run the model

loan_data

evaluate the result

training and test set



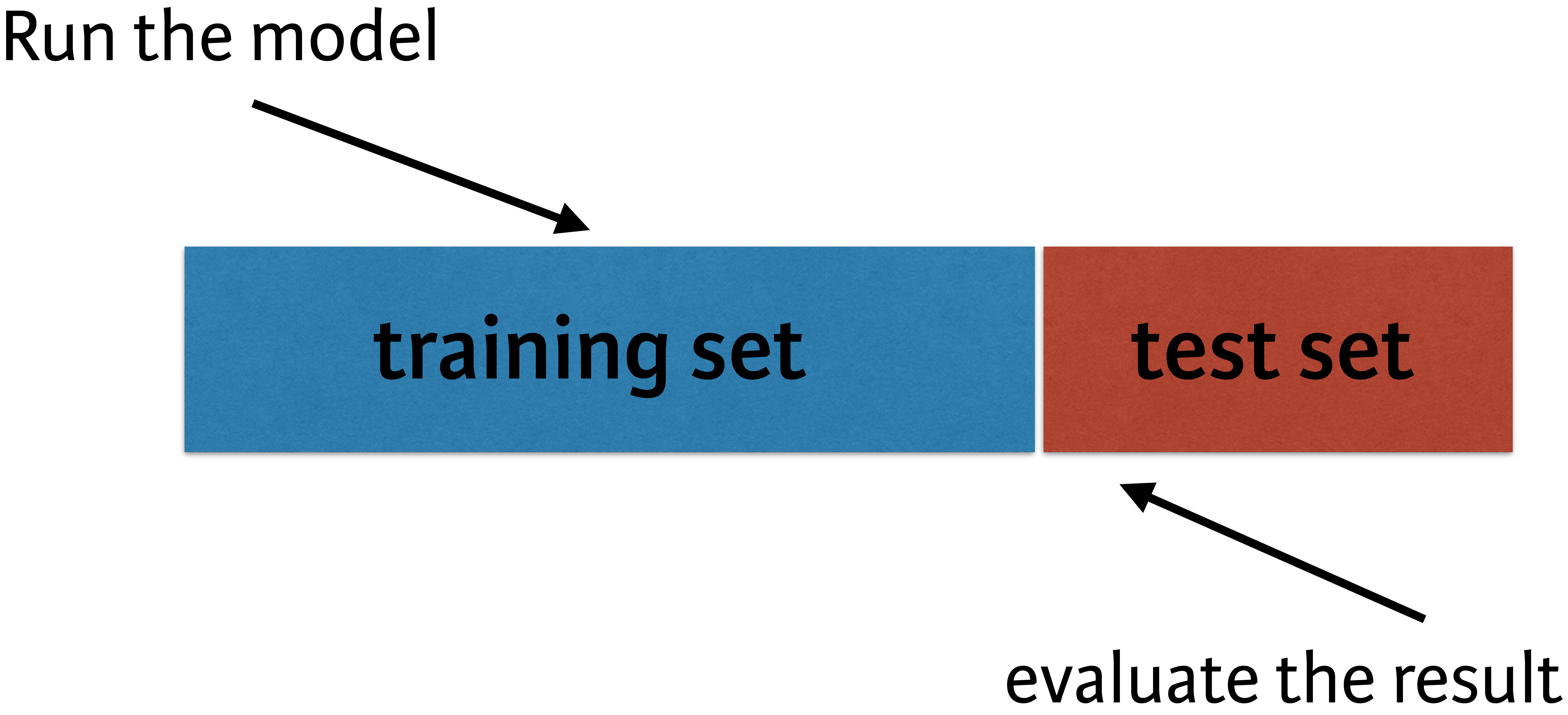
Run the model

training set

test set

evaluate the result

training and test set



Run the model

training set

test set

evaluate the result

Final data structure

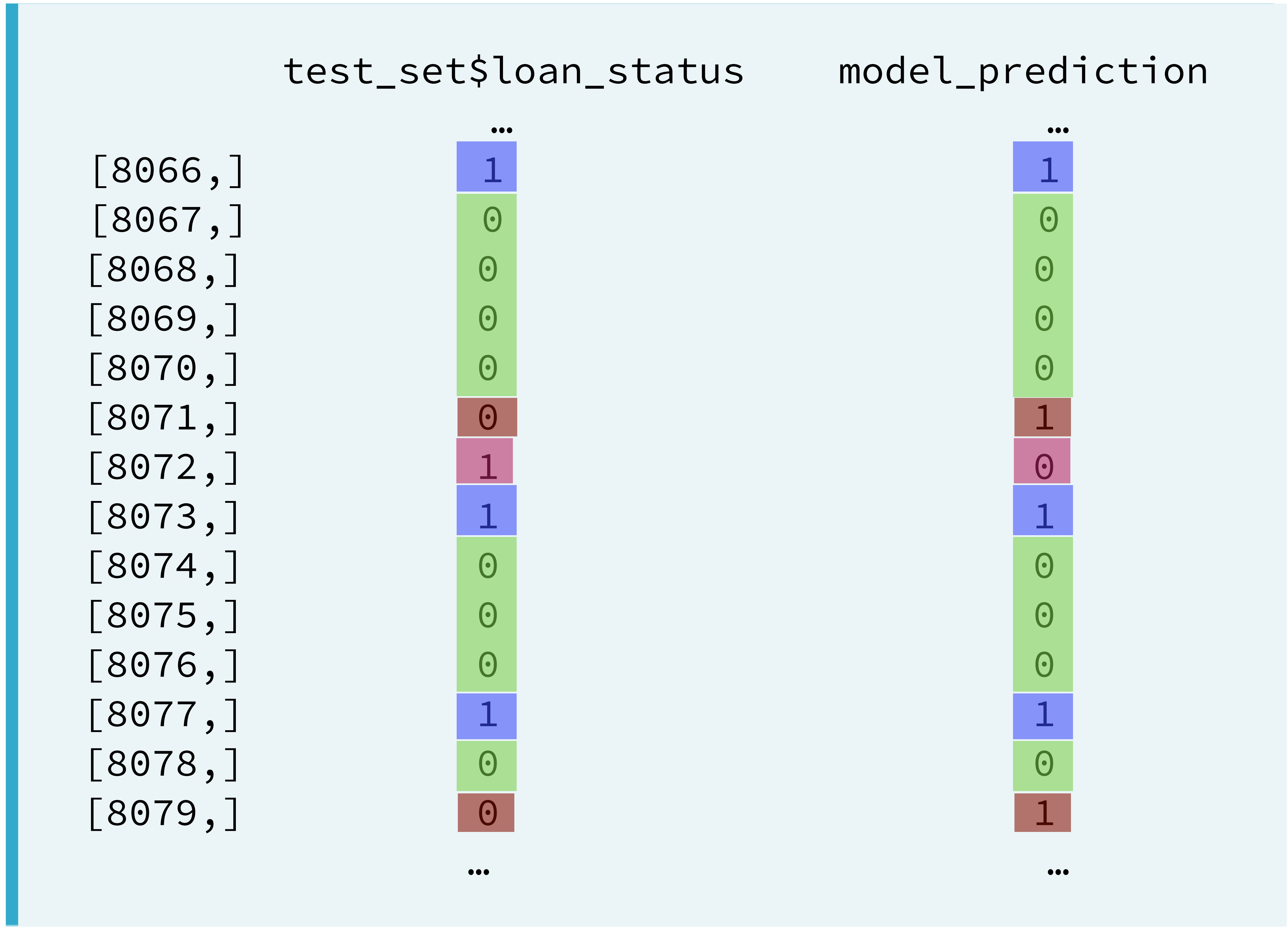
```
 head(training_set, 10)
         loan_status  loan_amnt  grade  home_ownership   annual_inc   age   emp_cat    ir_cat
21655              0      25000      B            RENT        91000    34      0-15   11-13.5
25468              0      16000      D            RENT        45000    25      0-15     13.5+
18407              0       8500      A        MORTGAGE       110000    29      0-15       0-8
14234              0       9800      B        MORTGAGE       102000    24      0-15      8-11
7588               0       3600      A        MORTGAGE        40000    59      0-15       0-8
7026               0       6600      A             OWN        26400    35     15-30       0-8
2180               0       3000      A            RENT        10000    24      0-15       0-8
14930              0       7500      B             OWN        27168    24      0-15      8-11
17083              0       6000      A            RENT        74970    26      0-15       0-8
15573              0      22750      A        MORTGAGE        32004    25      0-15       0-8
```

Final data structure

```
> str(training_set)

'data.frame': 19394 obs. of  8 variables:
 $ loan_status   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
 $ loan_amnt     : int  25000 16000 8500 9800 3600 6600 3000 7500 6000 22750 ...
 $ grade         : Factor w/ 7 levels "A","B","C","D",..: 2 4 1 2 1 1 1 2 1 1 ...
 $ home_ownership: Factor w/ 4 levels "MORTGAGE","OTHER",..: 4 4 1 1 1 3 4 3 4 1 ...
 $ annual_inc    : num  91000 45000 110000 102000 40000 ...
 $ age           : int  34 25 29 24 59 35 24 24 26 25 ...
 $ emp_cat       : Factor w/ 5 levels "0-15","15-30",..: 1 1 1 1 1 2 1 1 1 1 ...
 $ ir_cat        : Factor w/ 5 levels "0-8","11-13.5",..: 2 3 1 4 1 1 1 4 1 1 ...
```

evaluate a model



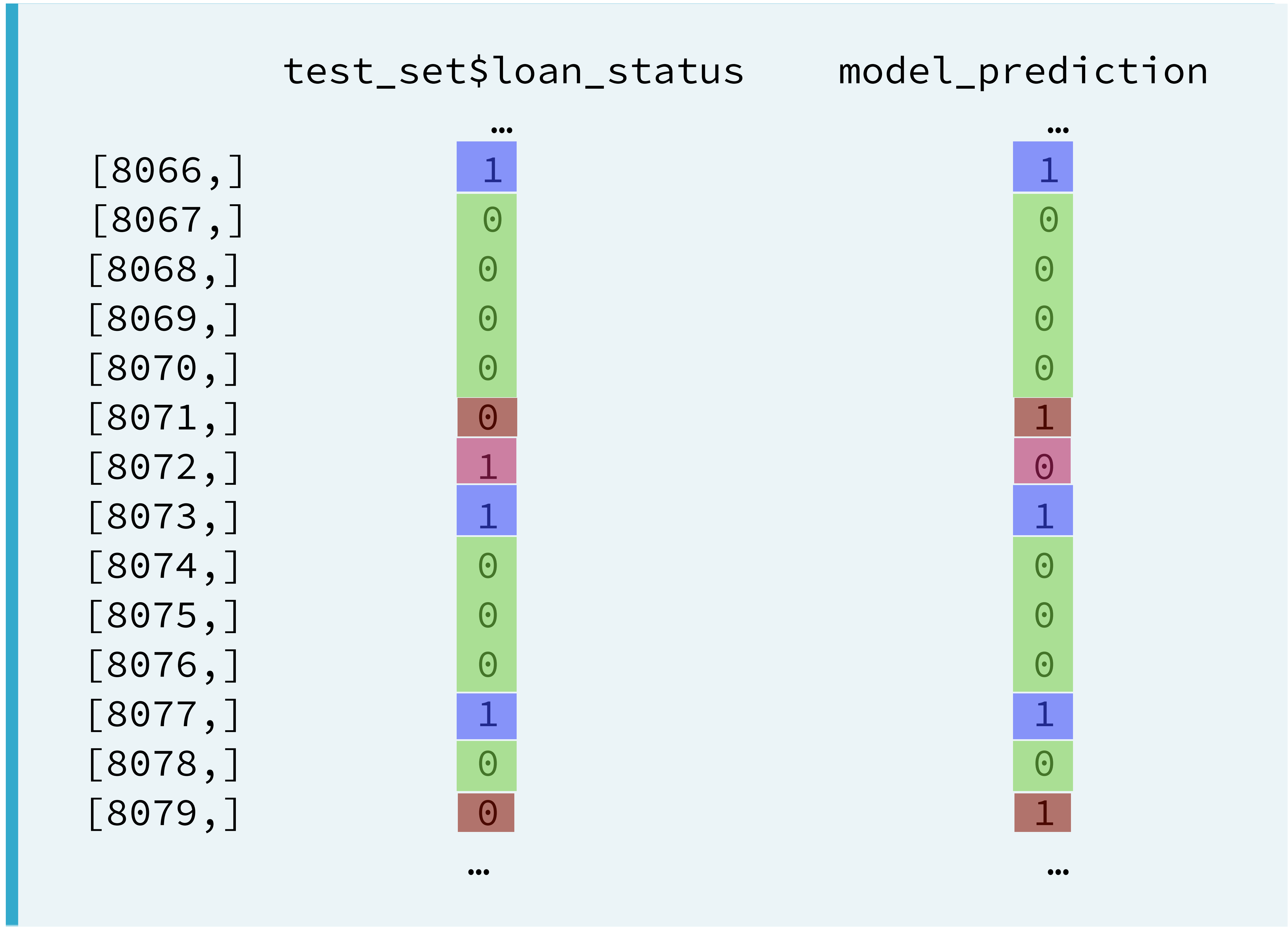
```
         test_set$loan_status    model_prediction
                  ...                  ...
[8066,]            1                    1
[8067,]            0                    0
[8068,]            0                    0
[8069,]            0                    0
[8070,]            0                    0
[8071,]            0                    1
[8072,]            1                    0
[8073,]            1                    1
[8074,]            0                    0
[8075,]            0                    0
[8076,]            0                    0
[8077,]            1                    1
[8078,]            0                    0
[8079,]            0                    1
                  ...                  ...
```

model prediction

actual loan status

	no default (0)	default (1)
no default (0)	8	2
default (1)	1	3

evaluate a model



```
         test_set$loan_status   model_prediction
                   ...                 ...
[8066,]             1                   1
[8067,]             0                   0
[8068,]             0                   0
[8069,]             0                   0
[8070,]             0                   0
[8071,]             0                   1
[8072,]             1                   0
[8073,]             1                   1
[8074,]             0                   0
[8075,]             0                   0
[8076,]             0                   0
[8077,]             1                   1
[8078,]             0                   0
[8079,]             0                   1
                   ...                 ...
```

model prediction

actual loan status

	no default (0)	default (1)
no default (0)	**TN**	**FP**
default (1)	**FN**	**TP**

some measures...

- Accuracy = (8 + 3) / 14 = 78.57%
- Sensitivity = 3 / (1 + 3) = 75 %
- Specificity = 8/ (8 + 2)= 80%

model prediction

actual loan status		no default (0)	default (1)
	no default (0)	8	2
	default (1)	1	3