
# Credit Risk Analysis
# Using Logistic Regression Modeling

**Introduction**

A loan officer at a bank wants to be able to identify characteristics that are indicative of people who are likely to default on loans, and then use those characteristics to discriminate between good and bad credit risks.

**Sample**

This case study uses information on 850 past and prospective customers to execute a **Logistic Regression Analysis**. Of these, 717 cases are customers who were previously given loans. We will use a random sample of 513 of these 717 customers to create a risk model. We will set aside the remaining 204 customers as a holdout or validation sample on which to test the credit-risk model; then use the model to classify the 133 prospective customers as good or bad credit risks. Binary logistic regression is an appropriate technique to use on these data because the "dependent" or criterion variable (the thing we want to predict) is dichotomous (loan default vs. no default).

First we display the crosstabulations below, which confirm our sample characteristics. The first table shows that we will be using 717 cases for building and validating our model, holding the 133 prospects aside for later scoring using the model's coefficients.

**Case Processing Summary**

| | Cases | | | | | |
|---|---|---|---|---|---|---|
| | Valid | | Missing | | Total | |
| | N | Percent | N | Percent | N | Percent |
| Previously defaulted * validate | 717 | 84.4% | 133 | 15.6% | 850 | 100.0% |

The second table shows that we have created a variable called "validate." Customers have been randomly assigned one of two values of this variable. The 513 customers who will be used to

build the model are assigned a value of 1. The remaining 204 customers will be assigned a value of zero, and will constitute the validation sample on which the model will be tested.

**Previously defaulted * validate Crosstabulation**

| | | | validate | | |
|---|---|---|---|---|---|
| | | | .00 | 1.00 | Total |
| Previously defaulted | No | Count | 163 | 410 | 573 |
| | | % within Previously defaulted | 28.4% | 71.6% | 100.0% |
| | | % within validate | 79.9% | 79.9% | 79.9% |
| | Yes | Count | 41 | 103 | 144 |
| | | % within Previously defaulted | 28.5% | 71.5% | 100.0% |
| | | % within validate | 20.1% | 20.1% | 20.1% |
| Total | | Count | 204 | 513 | 717 |
| | | % within Previously defaulted | 28.5% | 71.5% | 100.0% |
| | | % within validate | 100.0% | 100.0% | 100.0% |

The crosstabulations also show that the modeling sample contains 410 customers who did not default on a previous loan, and 103 who did default. The validation or holdout sample contains 163 customers who did not default, and 41 who did.

**Logistic Regression Analysis**

Now we will run a logistic regression modeling analysis and examine the results. Our model will be testing several candidate predictors, including:

- Age
- Level of education
- Number of years with current employer
- Number of years at current address
- Household income (in thousands)
- Debt-to-income ratio
- Amount of credit card debt (in thousands).

Our logistic regression modeling analysis will use an automatic stepwise procedure, which begins by selecting the strongest candidate predictor, then testing additional candidate predictors, one at a time, for inclusion in the model. At each step, we check to see whether a new candidate predictor will improve the model significantly. We also check to see whether, if the new predictor is included in the model, any other predictors already in the model should stay or be removed. If a newly entered predictor does a better job of explaining loan default behavior, then it is possible for a predictor already in the model to be removed from the model because it no longer uniquely explains enough. This stepwise procedure continues until all the candidate predictors have been thoroughly tested for inclusion and removal. When the analysis is finished,

we have the following table that contains various statistics.

For our purposes here, we can focus our attention on the "B" column, the "Sig." column and the "Exp(B)" column, explained below.

**Variables in the Equation**

| | | B | S.E. | Wald | df | Sig. | Exp(B) |
|---|---|---|---|---|---|---|---|
| Step 1[a] | debtinc | .369 | .037 | 100.408 | 1 | .000 | 1.446 |
| | Constant | -6.177 | .549 | 126.519 | 1 | .000 | .002 |
| Step 2[b] | debtinc | .334 | .038 | 75.260 | 1 | .000 | 1.396 |
| | creddebt | .329 | .085 | 15.164 | 1 | .000 | 1.390 |
| | Constant | -6.428 | .589 | 119.062 | 1 | .000 | .002 |
| Step 3[c] | employ | -.264 | .048 | 30.128 | 1 | .000 | .768 |
| | debtinc | .329 | .042 | 60.010 | 1 | .000 | 1.389 |
| | creddebt | .859 | .148 | 33.492 | 1 | .000 | 2.360 |
| | Constant | -5.632 | .635 | 78.761 | 1 | .000 | .004 |
| Step 4[d] | employ | -.262 | .050 | 28.020 | 1 | .000 | .769 |
| | address | -.078 | .033 | 5.639 | 1 | .018 | .925 |
| | debtinc | .327 | .042 | 59.154 | 1 | .000 | 1.386 |
| | creddebt | .949 | .162 | 34.242 | 1 | .000 | 2.584 |
| | Constant | -5.233 | .638 | 67.278 | 1 | .000 | .005 |

a. Variable(s) entered on step 1: debtinc.

b. Variable(s) entered on step 2: creddebt.

c. Variable(s) entered on step 3: employ.

d. Variable(s) entered on step 4: address.

The table's leftmost column shows that our stepwise model-building process included four steps. In the first step, a constant as well as the debt-to-income-ratio predictor variable ("debtinc") are entered into the model. At the second step the amount of credit card debt ("creddebt") is added to the model. The third step adds number of years at current employer ("employ"). And the final step adds number of years at current address ("address").

The "B" column shows the coefficients (called Beta Coefficients, abbreviated with a "B") associated with each predictor. We see that number of years at current employer and number of years at current address have negative coefficients, indicating that customers who have spent less time at either their current employer or their current address are somewhat more likely to default on a loan. The predictors measuring the debt-to-income ratio and amount of credit card debt both have positive coefficients, indicating that higher debt-to-income ratios or higher amounts of credit card debt are associated with a greater likelihood of defaulting on a loan.

The "Sig." column shows the levels of statistical significance associated with the various predictors in the model. The numbers essentially show us the likelihood that the predictor's

coefficient is spurious. The numbers are probabilities expressed as decimals. We want these numbers to be small, and they are, giving us our first indication that we appear to have a good model.

For example, the value of 0.018 associated with number of years at current address indicates that we would expect our model's result to deviate significantly from reality only about 18 times out of a thousand if we repeated our model-building process over and over again on new data samples. The statistical significance levels associated with the other three predictors are all smaller than 0.001 (one chance in a thousand of a spurious result), and so they are shown simply as 0.000.

While the "B" column is convenient for testing the usefulness of predictors, the "Exp(B)" column is easier to interpret. Exp(B) represents the ratio-change in the odds of the event of interest for a one-unit change in the predictor. For example, Exp(B) for number of years with current employer is equal to 0.769, which means that the odds of default for a person who has been employed at their current job for two years are just 0.769 times the odds of default for a person who has been employed at their current job for 1 year, all other things being equal.

Once a final model is created and validated, the information in the above table can be used to score the individual cases in a prospect database. This will allow the bank's marketing department to focus their acquisition efforts on those prospects that have the lowest model-predicted probability of defaulting on a loan. It is a simple matter to generate computer-readable instructions that can be used to quickly do the file scoring.

After building a logistic regression model, we need to determine whether it reasonably approximates the behavior of our data. There are usually several alternative models that pass the diagnostic checks, so we need tools to help us choose between them. Here are three types of tool that help ensure a valid model:

**Automated Variable Selection.** When constructing a model, we generally want to include only predictors that contribute significantly to the model. The modeling procedure that we used offers several methods for stepwise selection of the "best" predictors to include in the model, and we used one of these stepwise methods (Forward Selection [Likelihood Ratio]) to automatically identify our final set of predictors.

However, since we used a stepwise variable-selection procedure, the significance levels associated with the model predictors may be somewhat inaccurate because they are assuming a single-step process rather than a multi-step process. So we also use additional diagnostics to give us more confidence in our model.

**Pseudo R-Squared Statistics.** The well-known r-squared statistic, which measures the variability in the dependent variable that is explained by a linear regression model, cannot be computed for logistic regression models because our dependent variable is dichotomous rather

than continuous. So we instead use what are called pseudo r-squared statistics. The pseudo r-squared statistics are designed to have similar properties to the true r-squared statistic. The table below shows that as our stepwise procedure moved forward from step one to step four, the pseudo r-squared statistics became progressively stronger. For those who are familiar with the r-squared statistic from linear regression, the Nagelkerke statistic in the far righthand column represents a good approximation to that statistic, having a maximum possible value of 1.00. It shows that approximately 72% of the variation in the dependent variable is explained by the four predictors in our final model.

**Model Summary**

| Step | -2 Log likelihood | Cox & Snell R Square | Nagelkerke R Square |
|---|---|---|---|
| 1 | 269.081[a] | .380 | .601 |
| 2 | 251.889[b] | .401 | .633 |
| 3 | 206.451[b] | .451 | .713 |
| 4 | 200.235[b] | .458 | .723 |

a. Estimation terminated at iteration number 6 because parameter estimates changed by less than .001.

b. Estimation terminated at iteration number 7 because parameter estimates changed by less than .001.

We also test the model's goodness-of-fit using additional diagnostics (not shown here), and these additional tests also confirm that we have a good model.

**Classification and Validation.** Crosstabulating observed response categories with predicted categories helps us to determine how well the model identifies defaulters. Here is a classification and validation crosstabulation table:

**Classification Table[d]**

| | | | Predicted | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | Selected Cases[a] | | | Unselected Cases[b,c] | | |
| | Observed | | Previously defaulted | | Percentage Correct | Previously defaulted | | Percentage Correct |
| | | | No | Yes | | No | Yes | |
| Step 1 | Previously defaulted | No | 385 | 25 | 93.9 | 149 | 14 | 91.4 |
| | | Yes | 47 | 56 | 54.4 | 13 | 28 | 68.3 |
| | Overall Percentage | | | | 86.0 | | | 86.8 |
| Step 2 | Previously defaulted | No | 388 | 22 | 94.6 | 151 | 12 | 92.6 |
| | | Yes | 43 | 60 | 58.3 | 11 | 30 | 73.2 |
| | Overall Percentage | | | | 87.3 | | | 88.7 |
| Step 3 | Previously defaulted | No | 395 | 15 | 96.3 | 153 | 10 | 93.9 |
| | | Yes | 28 | 75 | 72.8 | 7 | 34 | 82.9 |
| | Overall Percentage | | | | 91.6 | | | 91.7 |
| Step 4 | Previously defaulted | No | 395 | 15 | 96.3 | 154 | 9 | 94.5 |
| | | Yes | 25 | 78 | 75.7 | 8 | 33 | 80.5 |
| | Overall Percentage | | | | 92.2 | | | 91.7 |

a. Selected cases validate EQ 1

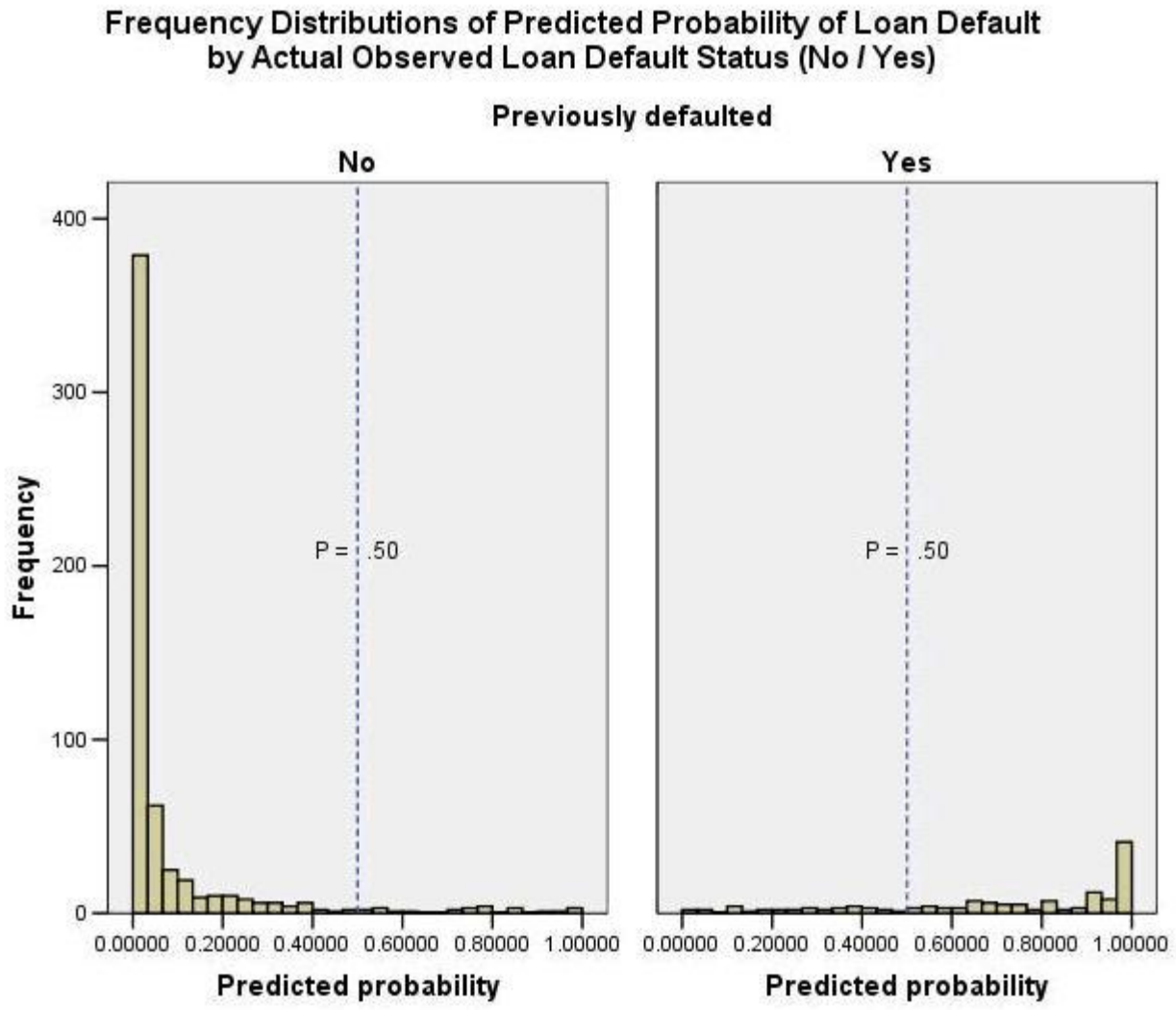
b. Unselected cases validate NE 1

c. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

d. The cut value is .500

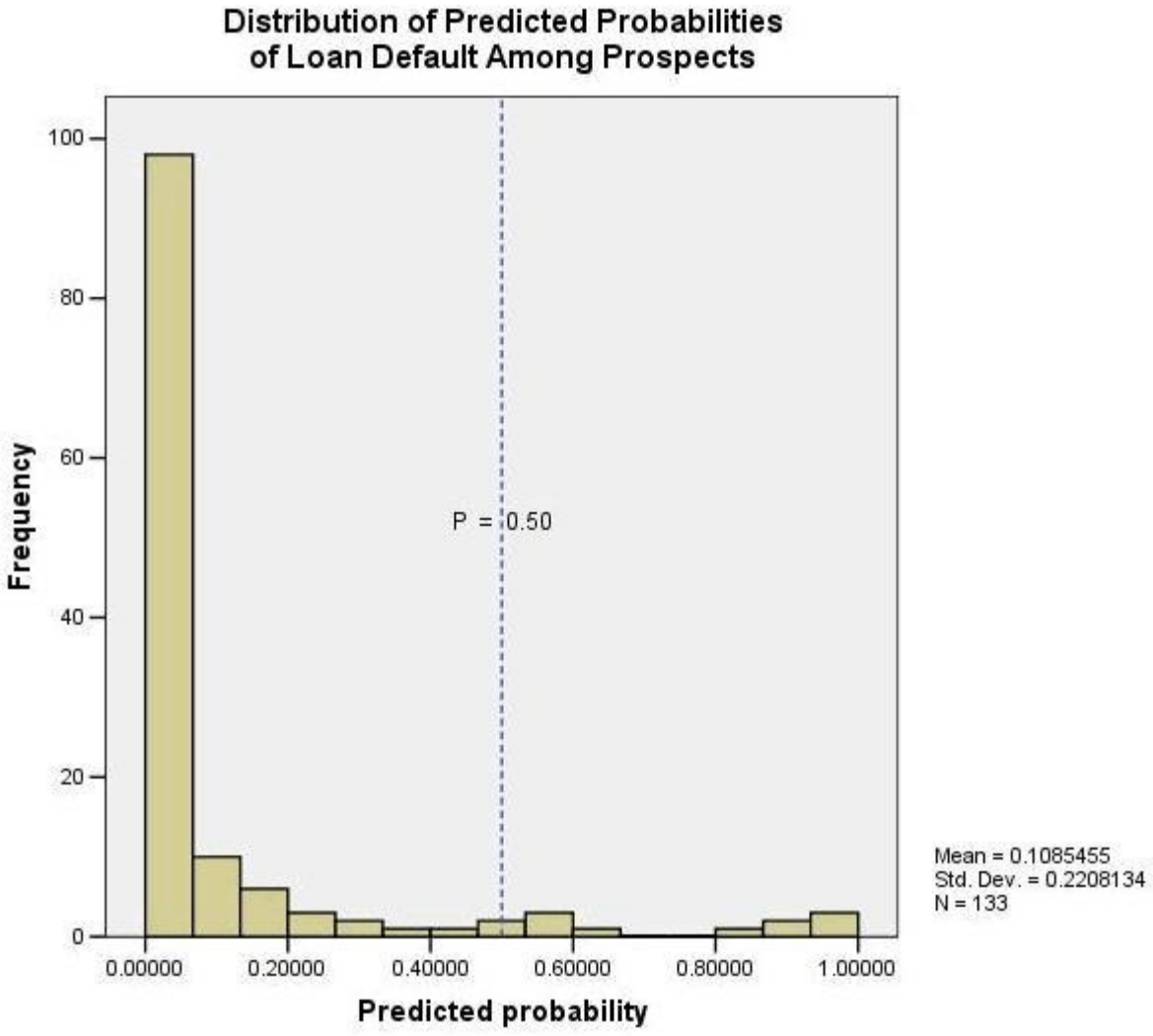
The table shows that the model correctly classified about 96% of the modeling sample's non-defaulters and about 76% of the modeling sample's defaulters, for an overall correct classification percentage of about 92%. Similarly, when applied to the holdout or validation sample, the model correctly identified about 95% of the non-defaulters and about 81% of the defaulters, for an overall correct classification percentage of about 92%.

The double-panel graph below provides additional information about the model's strength. It shows probability distributions for the probability of defaulting, separately for actual non-defaulters and actual defaulters. The binary logistic regression model assigns probabilities of defaulting to each customer, ranging from zero to 1.00 (zero to 100%). In this case, it uses a cut point of exactly 0.50 (50% probability) as the dividing line between predicted non-defaulters and predicted defaulters.




The leftmost graph shows that the modeling process assigned the bulk of the actual non-defaulters very low probabilities of defaulting, far below the 50% probability cut point. And the righthand graph shows that the model assigned the bulk of the defaulters very high probabilities of defaulting, far above the 50% cut point. So this adds more confirmation that we have a good model.

Now that we have a valid predictive model, we can use it to score a prospect file. The graph below shows the result after we have scored our 133 prospects.




It shows that approximately 67% of the prospects would not be expected to default on a loan. (If we had used much larger customer and prospect samples, as would typically be the case, then the prospect sample's results would more closely resemble the modeling sample's results.) Note that the separation of prospects into predicted defaulter and non-defaulter subgroups is not quite as clean as for the modeling sample. Although larger samples would mitigate this difference, it is typical for a model to deteriorate slightly when applied to a sample that is different from the one on which the model was built, due to natural sampling error.

A critical issue for loan officers is the cost of what statisticians refer to as Type I and Type II errors. That is, what is the cost of classifying a defaulter as a non-defaulter (Type I error)? And what is the cost of classifying a non-defaulter as a defaulter (Type II error)?

If bad debt is the primary concern, then we want to lower our Type I error and maximize our "sensitivity". (Sensitivity is the probability that a "positive" case [a defaulter] is correctly

classified.) If growing our customer base is the priority, then we want to lower our Type II error and maximize our "specificity". (Specificity is the probability that a "negative" case [a non-defaulter] is correctly classified.)

Usually both are major concerns, so we have to choose a decision rule for classifying customers that gives the best mix of sensitivity and specificity. In our example we arbitrarily chose a probability cut point of 0.50 (50%). But in practice, depending on our specific objectives, we may want to experiment with various cut points to see how these affect our models' sensitivity and specificity by examining the rates of correct classification for each model.

**Summary**

We have demonstrated the use of **risk modeling** using **logistic regression analysis** to identify demographic and behavioral characteristics associated with likelihood to default on a bank loan. We identified four important influences, and we confirmed the validity of the model using several diagnostic analytic procedures. We also used the results of the model to score a prospect sample, and we briefly discussed the importance of examining a model's sensitivity and specificity in the context of one's specific, real-world objectives.

---

The foregoing case study is an edited version of one originally furnished by SPSS, Inc., and is used with their permission.