Logistic Regression - Analysis

Introduction

We're going to look at the same data set from Lending Club but ask a different question. One that has a binary outcome. Let's assume we have a FICO Score of 720 and we want to borrow 10,000 dollars. We would like to get an Interest Rate less that 12 per cent. The question we pose is: ## Can we get a loan, from the Lending Club, of 10,000 dollars at 12 per cent or less, with a FICO Score of 720?

Methods

How do we use Logistic Regression here? Let's recast the problem as follows:-

What is the probability of getting a Loan, from the Lending Club, of 10,000 dollars at 12 per cent or less with a FICO Score of 720?

Then let us decide that if we get a probability of less than 0.67 we say it means we won't get the loan and if it is greater than 0.67 we will. I.e. we are not confident until we have a 2/3 chance of getting it.

In reality we can set the threshold higher, say 0.8, if we want to be "more certain" that it will happen, but for this exercise we'll just say 0.67.

From initial discussion we say we want to start with a model of the form

$InterestRate = a_0 + a_1 * FICOScore + a_2 * LoanAmount$

And the derive a second equation of the form:

Z = Prob (InterestRate less than 12 percent).

We apply this to the existing dataset and create a Logistic Regression Model using modeling software.

Results

As with the Linear Regression Model, we use the cleaned up Lending Club data set as input.

```
pander(dfr[1:5,])
```

Interest.Rate	FICO.Score	Loan.Length	Monthly.Income	Loan.Amount
15.31	670	36	4892	6000
19.72	670	36	3575	2000
14.27	665	36	4250	10625
21.67	670	60	14167	28000
21.98	665	36	6667	22000

```
## we add a column which indicates (True/False) whether the interest rate is <= 12
dfr['TF'] = dfr['Interest.Rate']<=12
# inspect again
pander(dfr[1:5,])
```

Interest.Rate	FICO.Score	Loan.Length	Monthly.Income	Loan.Amount	TF
15.31	670	36	4892	6000	FALSE
19.72	670	36	3575	2000	FALSE
14.27	665	36	4250	10625	FALSE
21.67	670	60	14167	28000	FALSE
21.98	665	36	6667	22000	FALSE

```
# we see that the TF values are False as Interest.Rate is higher than 12 in all these cases
```

```
# now we check the rows that have interest rate == 10 (just some number < 12)
# this is just to confirm that the TF value is True where we expect it to be
d = dfr[dfr$Interest.Rate == 10,]
pander(d[1:5,])
```

Table 3: Table continues below

	Interest.Rate	FICO.Score	Loan.Length	Monthly.Income
1057	10	700	36	3250
1526	10	715	36	15417
1586	10	730	36	6250
1607	10	715	36	5000
1722	10	735	60	4000

	Loan.Amount	TF
1057	2800	TRUE
1526	6000	TRUE
1586	21000	TRUE
1607	12000	TRUE
1722	5000	TRUE

```
# all is well
```

Now we use our Logistic Regression modeler software to create Logit model using this data, with the 'TF' column as the dependent (or response) variable and 'FICO.Score' and 'Loan.Amount' as independent (or predictor) variables.

```
#Fitting Logistic Regression
#Using the logit model
#The code below estimates a logistic regression model using the glm (generalized linear model) function
logit <- glm(TF ~ FICO.Score + Loan.Amount, data = dfr, family = "binomial")
```

We should see some soothing messages from our software re-assuring us that all went well and giving us some numbers we may not find useful right now. More importantly we want the results. What are the fitted coefficients that the software has computed?

```
coeff = logit$coefficients
coeff
```

```
##   (Intercept)     FICO.Score   Loan.Amount
## -60.125045279   0.087423216  -0.000174028
```

```r
# get the fitted coefficients from the results
summary(logit)
```

```
##
## Call:
## glm(formula = TF ~ FICO.Score + Loan.Amount, family = "binomial",
##     data = dfr)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.5099  -0.4694  -0.1764   0.3526   2.9082
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -60.125045   2.420450  -24.84   <2e-16 ***
## FICO.Score    0.087423   0.003528   24.78   <2e-16 ***
## Loan.Amount  -0.000174   0.000011  -15.81   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3358.5  on 2499  degrees of freedom
## Residual deviance: 1597.5  on 2497  degrees of freedom
## AIC: 1603.5
##
## Number of Fisher Scoring iterations: 6
```

The numbers above are the coefficients for the respective independent, i.e. predictor, variables in the linear expression we saw in the Overview. Except, we now have two instead of one predictor. We have multivariate linear regression.

So, using the above coefficients, the linear part of our predictor is

$$z = -60.125 + 0.087423 * FicoScore - 0.000174 * LoanAmount$$

Finally, the probability of our desired outcome, ie our getting a loan at 12% interest or less, is

$$p(z) = \frac{1}{1 + e^{-}(b_0 + b_1 * FicoScore + b_2 * LoanAmount)}$$

where b_0 = -60.125, b_1 = -0.087423 and b_2 = -0.000174

We create a function in code that encapsulates all this.

It takes as input, a borrowers FICO score, the desired loan amount and the coefficient vector from our model. It returns a probability of getting the loan, a number between 0 and 1.

```r
pz <- function(fico,amt,coeff) {
  # compute the linear expression by multipyling the inputs by their respective coefficients.
  z = coeff[1] + coeff[2]*fico + coeff[3]*amt
  return(1/(1+exp(-z)))
  }
```

Now we use our data FICO=720 and Loan Amount=10,000 to get a probability using the z value and the logistic formula.

```
pz(720,10000,coeff)
```

```
## (Intercept)
##   0.7463786
```

This value of 0.746 tells us we have a good chance of getting the loan we want, according to our criterion, where anything above 0.67 was considered a 'yes'.

Now we are going to try (fico, amt) pairs as follows:

- 720,20000
- 720,30000
- 820,10000
- 820,20000
- 820,30000

```
print("Trying multiple FICO Loan Amount combinations: ")
```

```
## [1] "Trying multiple FICO Loan Amount combinations: "
```

```
print('----')
```

```
## [1] "----"
```

```
print("fico=720, amt=10,000")
```

```
## [1] "fico=720, amt=10,000"
```

```
print(pz(720,10000,coeff))
```

```
## (Intercept)
##   0.7463786
```

```
print("fico=720, amt=20,000")
```

```
## [1] "fico=720, amt=20,000"
```

```
print(pz(720,20000,coeff))
```

```
## (Intercept)
##   0.3405399
```

```
print("fico=720, amt=30,000")
```

```
## [1] "fico=720, amt=30,000"
```

```
print(pz(720,30000,coeff))
```

```
## (Intercept)
##   0.0830836
```

```
print("fico=820, amt=10,000")
```

```
## [1] "fico=820, amt=10,000"
```

```
print(pz(820,10000,coeff))
```

```
## (Intercept)
##   0.9999457
```

```
print("fico=820, amt=20,000")
```

```
## [1] "fico=820, amt=20,000"
```

```
print(pz(820,20000,coeff))
```

```
## (Intercept)
##   0.9996909
```

```
print("fico=820, amt=30,000")
```

```
## [1] "fico=820, amt=30,000"
```

```
print(pz(820,30000,coeff))
```

```
## (Intercept)
##   0.9982408
```

We see as somewhat expected that the person with a 720 FICO Score will have decreasing probability of getting loans with higher amounts. However, the person with the 820 FICO Score is very likely to get loans with those amounts, again as expected.

```
pz(820,63000,coeff)
```

```
## (Intercept)
##   0.6452512
```

Exercise

Try the following pairs of (fico, amt) values and plug them into the pz() function mimicing the syntax below. What insight does this give you?

- 820,50000
- 820,60000
- 820,70000
- 820,63000
- 820,65000

Place your cursor on the cell below. Hit shift-enter to recreate the result. Then click Insert->Cell Below via the Insert menu dropdown. This creates a new empty cell. Now enter the pz() function with the next pair of values. Hit shift-enter. Repeat this till the end of the list of values. Answer the question above, if possible. Then explore other pairs as you wish.

```
pz(820,50000,coeff)
```

```
## (Intercept)
##   0.9458637
```

Challenge Exercise

Use the supporting notebooks in the appendix to learn some plotting techniques and try to create a yes/no plot for loan amount on x-axis and probability of loan on the y-axis for a FICO score of 720. Do the same for a fico score of 820.

Extra Challenge Exercise

How would you create a plot that showed the probability of getting a loan as a function of both FICO score and loan amount varying? What tools would you need?

Conclusion

We see for the (720, 10000) case, a probability close to 0.7 which tells us that we have a good chance of getting the loan at a favorable interest rate. Using our threshold of 0.67 we count this as a 'yes'.

Using a Logistic Regression model, a desired Interest Rate of 12 per cent, we use dthe Lending Club dataset to compute a probability that we will get a 10,000 dollar loan with a FICO Score of 720. Our result indicated with a strong degree of certainty that we would be able to procure a loan with these terms.

When we try the multiple combinations we see the following:

- With a FICO Score of 720 the chance of a 20,000 and 30,000 Loan is lower than 0.67 so we count that as a probable "no".
- For the same amounts the FICO=820 score corresponds to probabilities greater than 0.75 and we count that as a "yes".
- For the same FICO the probability goes down with increasing Loan Amount
- For the same Loan Amount, the lower FICO has a lower probability.
- This is consistent with the signs of the coefficients for these variables in our model.