Credit Modelling

Chris Hua
Kevin Huo
Arjun Jain
Juan Manubens

11/7/2016

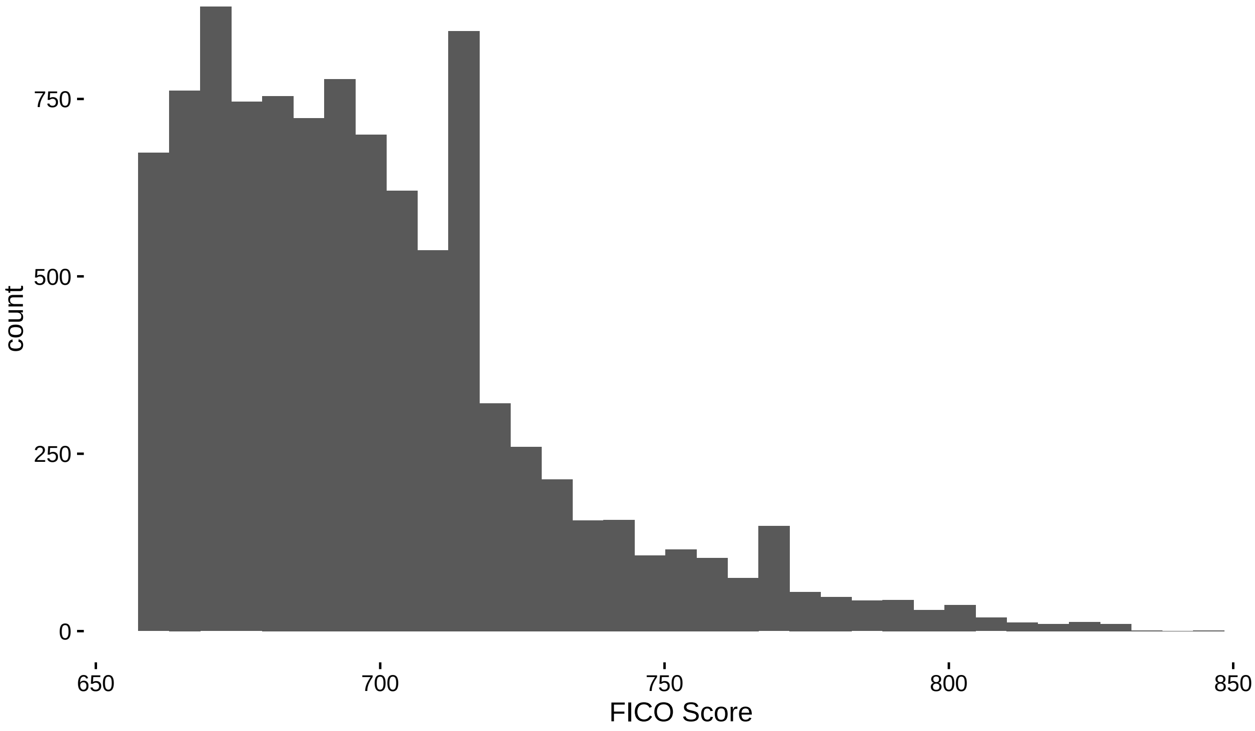
forgive me

1: Basic Modelling

We model "default probability" with respect to "FICO" score only. First, we note the distribution of the outcome variable:

```
.
Defaulted      Paid
     1623      8377
```

As well as the distribution of the FICO score, among applicants in this dataset. This is a pretty heavily left-skewed distribution, which probably represents the fact that the people who try to get loans on Lending Club are people who can't get loans traditionally.



a. We expect a negative coefficient on FICO. Intuitively, a customer is less likely to default the better their credit score is. We would also expect a positive intercept, because for a customer with 0 credit score (the condition for the intercept), they have terrible chances of paying off the bill, and thus high chance for default.

b. We estimate our model using `glm` approach with a binomial prior. Summary statistics:

term	estimate	std.error	statistic	p.value
(Intercept)	-7.68	0.76	-10.17	0
fico	0.01	0.00	12.29	0

By p-value, `fico` is a statistically significant value at better than at 0.01 level, indicating that by itself, `fico` is a useful predictor for default outcomes.

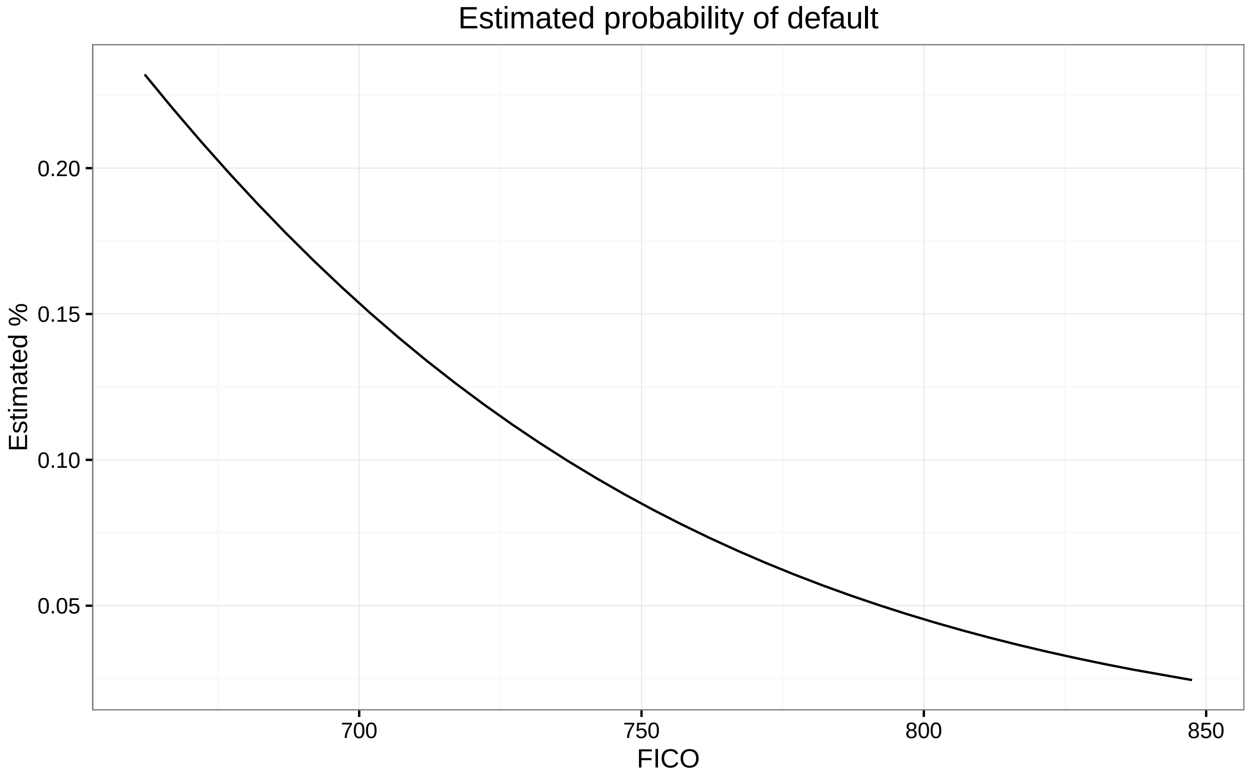
The intercept is strongly positive, which indicates that default is very likely for somebody with a `fico` score of 0. Alternatively, the intercept suggests that somebody with a score of 768 has 0 chance of default. The positive coefficient means that for each point of FICO score, the likelihood of a negative outcome decreases by 0.01. Each of these observations lines up with our intuitions.

2: Model Evaluation

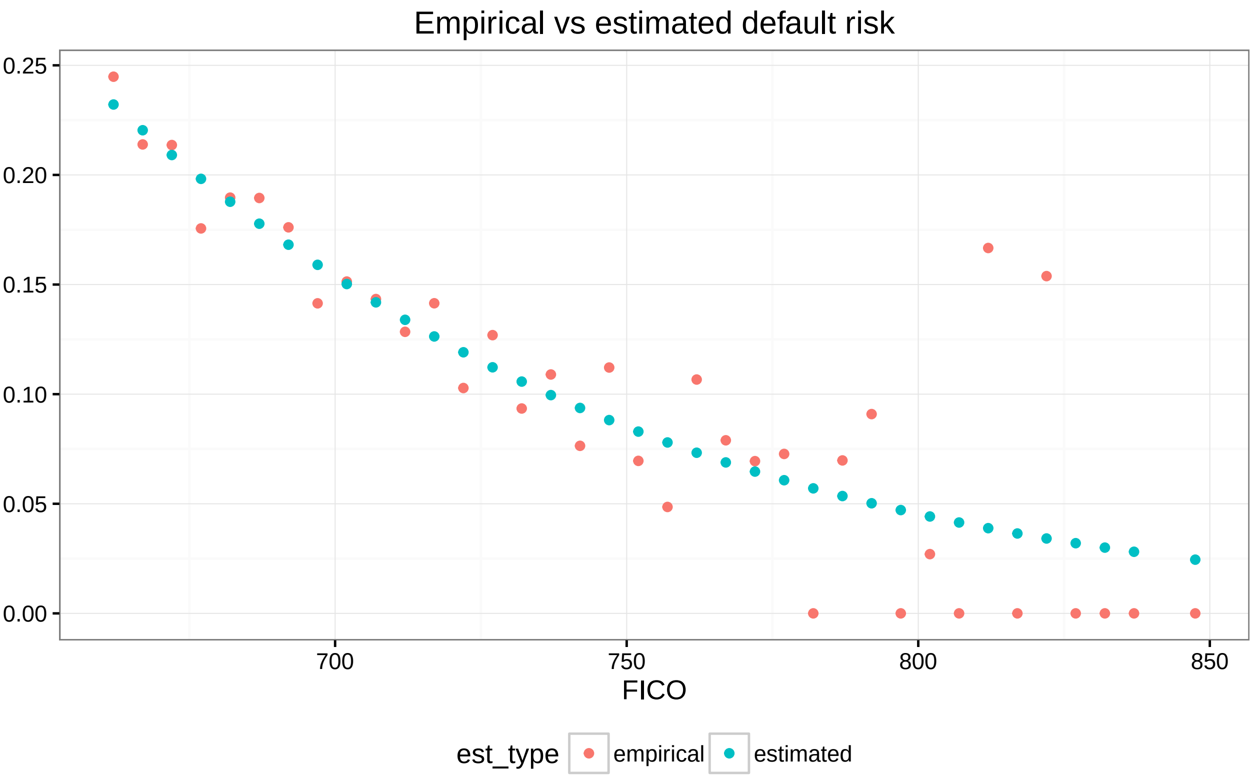
(a) We can estimate a probability of default via the formula from class 4:

$$Pr(f) = \frac{\exp(\beta_0 - \beta_1 \times f)}{1 + \exp(\beta_0 - \beta_1 \times f)}$$

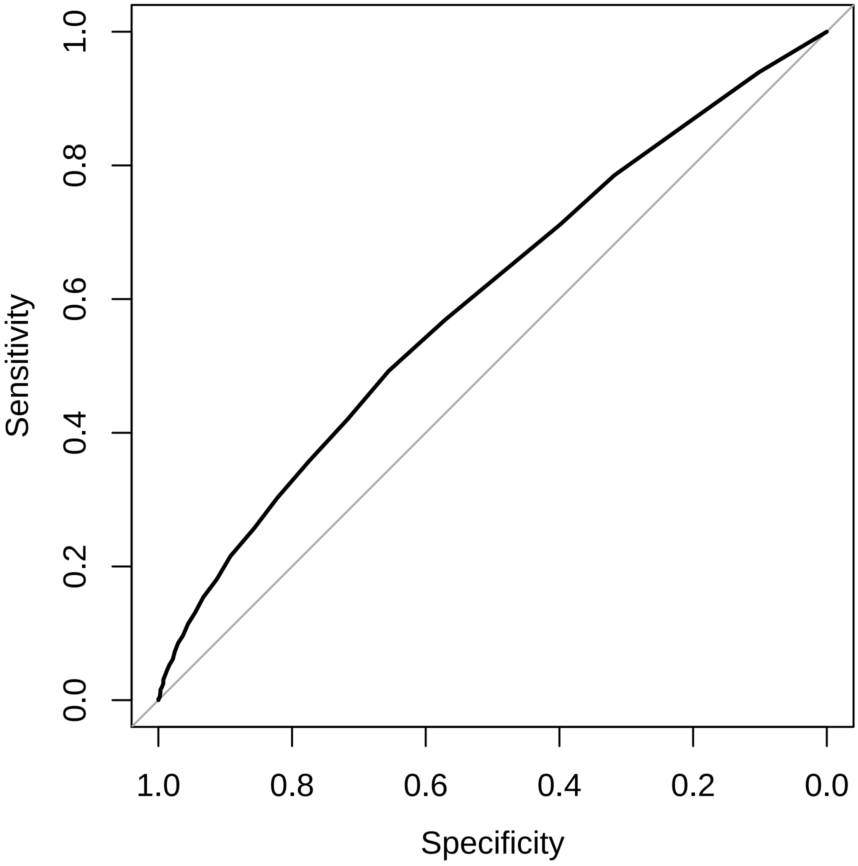
Since our model only includes FICO scores, we can easily plot FICO vs estimated probability of default.



Interestingly, we can compare this plot to the empirically determined default risk. We can see that the estimated curve fits very well, except for the outliers in high credit score. This might tell us that high credit score borrowers have asymmetric information about their ability to pay off loans, and they are the ‘lemons’ in this market.



(b) Then, we can plot a corresponding ROC curve for this model:



```
##
## Call:
## roc.formula(formula = default ~ prob, data = .)
##
## Data: prob in 1623 controls (default Defaulted) > 8377 cases (default Paid).
## Area under the curve: 0.5981
```

(c) For this model, we have AUC of 0.5981371. This is better than 0.5. It's consistent with the findings in part 1b, where we found that FICO was a statistically significant predictor alone of default.

(d) This is essentially using a 0.1 threshold on our probabilities of default. Then, we can create a confusion matrix:

	Defaulted	Paid
FALSE	89	1095
TRUE	1534	7282

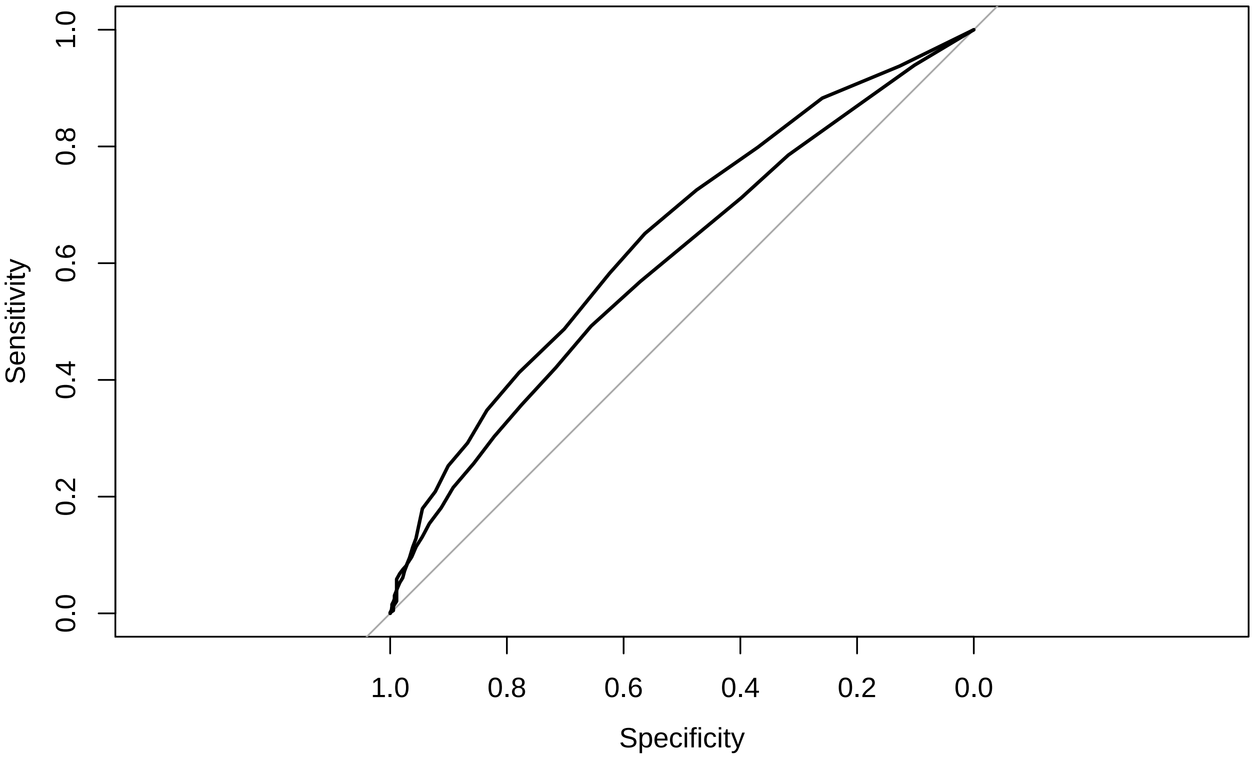
The proportion correctly rejected is 0.0548367. The proportion mistakenly rejected is 0.9451633. This is no bueno - our threshold is probably too low. # 3: An out-of-sample analysis

(a) The new model, estimated on the first 9000 rows, is given by these summary statistics:

term	estimate	std.error	statistic	p.value
(Intercept)	-7.16	0.79	-9.06	0
fico	0.01	0.00	11.11	0

(b) Then, we can predict probabilities for the remaining loans, and then create an ROC curve for both fits:

```
##
## Call:
## roc.formula(formula = default ~ prob, data = .)
##
## Data: prob in 1623 controls (default Defaulted) > 8377 cases (default Paid).
## Area under the curve: 0.5981
```



```
##
## Call:
## roc.formula(formula = default ~ pred, data = ., )
##
## Data: pred in 181 controls (default Defaulted) > 819 cases (default Paid).
## Area under the curve: 0.6459
```

(c) The area below the new ROC curve gets larger. We are no longer overfitting our dataset.

(d) Vacuously, you don't want to use all variables available, because some of the variables are unique to a person or a loan- e.g. "id". These could perfectly estimate somebody's probability of default in the sample set but have no predictive use.

[overfit…]

Finally, we also may want parsimoniousness […]

e Let's consider interest rate, loan length, and annual income. We're going to do this on test-train split as well.

We fit the model and show the Anova table (type II tests):

```
## Analysis of Deviance Table (Type II tests)
##
## Response: default
##            LR Chisq Df Pr(>Chisq)
## fico         5.8760  1    0.01535 *
## int_rate     7.8866  1    0.00498 **
## emp_length   7.2645 10    0.70026
## annual_inc   9.2386  1    0.00237 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Within this model, each variable is significant at the 0.05 level except the length of the loan. We can kick out that variable to create a more parsimonious, 3 variable, model.

```
## Analysis of Deviance Table (Type II tests)
##
## Response: default
##            LR Chisq Df Pr(>Chisq)
## fico         7.1142  1  0.0076476 **
## int_rate     8.2443  1  0.0040880 **
## annual_inc  11.1725  1  0.0008302 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can test the accuracy of this model

```
##
## Call:
## roc.formula(formula = default ~ pred, data = .)
##
## Data: pred in 181 controls (default Defaulted) < 819 cases (default Paid).
## Area under the curve: 0.6721
```

4: A business decision to make

```
## Warning: NAs introduced by coercion
```

Since we must pick 100 loans to finance, we need to construct a benchmark return on a low risk loan judging from the available options.

Our expected value is, for any given loan i,:

$$EV_i = \frac{1-(1+r_i)^{-n}}{r_i} \times \Pr(payoff \mid X_i) \times A_i$$

Then, $\frac{1-(1+r_i)^{-n}}{r_i}$ is a standard present value of annuity formula, where r_i is the interest rate we charge and n is the number of years that the loan will run. Finally, $\Pr(payoff \mid X_i)$ is the probability that the user will pay off the loan, given some characteristics X_i at the user level, and A_i is the amount of the loan. Note that the payoff probability comes from the model which we estimate in part 3.

This is a reasonable function because we care about the interest to be earned over the life of the loan, the chance of default, and the total amount of the loan. This is simplified somewhat because we have both liquidity constraints and investment constraints - in the absence of investment constraints we would allocate some money to this risky loan portfolio but also to a risk-free or market portfolio, and in the absence of liquidity constraints we would back each loan with a positive expected value. We additionally assume that loans are either fully paid off or fully written off, which isn't a reasonable assumption in the real world, since we can usually expect to recoup some of the value of a failed loan.

On these 100 loans, we expect a 0.0445021 return. That sucks.