Credit Risk Modeling for Online Consumer Loans

Matthew Dixon & Litong Dong
University of San Francisco

May 26, 2015

1 Executive summary

Institutional investors and investment managers seek to better characterize the credit risk of online consumer loans. This article describes how to prepare the data and build a credit risk model that can be used for a number of applications including generating alpha, issuing protection and securitizing loans into bonds with the desired risk/reward profile. A simple example is used to provide insight into the modeling approach. In the next blog article, we shall describe how to price consumer loans and include pre-prepayment models.

2 Background

Institutional investors and investment managers seek to better characterize the credit risk by not only using past performance of the loan but also additional information about the borrower. More accurate characterization of the credit risk leads to a reduction in the loan valuation model risk. Better credit risk models can be used for a variety of applications including

- building securitized products with the desired risk/reward profile;
- generate alpha by purchasing securities the investor perceives as being underpriced; and
- evaluating the cost of protection in the form of credit derivatives.

All of these factors broaden the accessibility of this asset class to institutional investors and investment managements.

While the financial engineering products emerging in this space mirror those seen in fixed income, credit risk modeling for peer-to-peer loans has a number of distinct features. First, consumer loans are illiquid due to a thin secondary market and hence it is not possible to build a credit model from market closing prices.

Second, peer-to-peer lending companies such as Prosper and Lending Club provide a rich set of information on the profile of the borrower. So what one loses from unavailability of daily updated data, one gains from the profile characteristics of the borrower. Building a credit risk model tailored to the data turns out to be the key to building an effective credit risk model.

2.1 Literature review

Many consumer loan modeling studies [2][3] have focussed on predicting default as a binary outcome without modeling time to default. Being able to model time to default, however, is crucial for modeling the fair value of the loan since this depends on the estimate of the likely number of loan installments that will be made. Furthermore, structured products and related credit derivatives rely on a model for time to default.
This article provides an overview of a general methodology for estimating the probability of default over the life of the loan that is agnostic to the loan platform. An important take-away from this article is the description of how the survival data is prepared. This preparation stage requires many assumptions and is often not made transparent by other studies which are methodology centric.

3 Preparing the data

Using the quarterly published loan data from Lending Club (as of January 1st, 2015), we created a survival dataset by finding the last recorded status and computing the survival time for each loan. First we defined the timeline in Figure 1 that maps known variables to new variables. By using the known variables ”issue date” and ”payment due date”, we compute the duration of the loan by finding their difference. Then we let the day that is 122 days after the last payment date be our default day (which is 91 days later than the virtual payment due date). Commensurate with that definition, we assign the status of each loan as either ”delinquent” or ”active”.

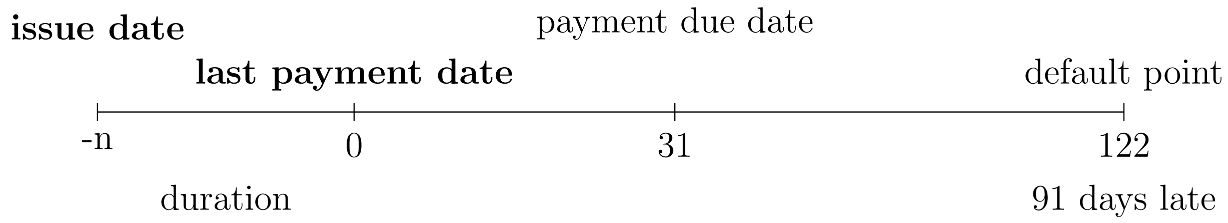

Figure 1: Timeline

In summary we applied the following mappings:

{“Current”, “Fully Paid”} →“active”
{“Charged Off”, “Default”} → “deliquent”

The remaining three statuses ”In Grace Period”, ”Late (16-30 days)” and ”Late (31-120 days)” required more refined definitions. We checked the number of days between their last payment dates to the date on which the data was pulled (virtual “today”), if the number is greater than 122, we considered the loan to be “delinquent”; otherwise, the loan was considered “active”.

4 Model

Following the approach described by [1], this research applies survival analysis to characterize the probability of default over the lifetime of a loan. Cox's method of proportional hazards splits the survival functions into two components: an underlying, baseline hazard function and an effects parameter describing the effect of explanatory variables on the loan performance. The model builder has some degree of flexibility in choosing these explanatory variables from a large number of borrower characteristics, but it is essential that the fitted coefficients are statistically significant otherwise such variables are somewhat meaningless.
The baseline hazard function is the cumulative probability of default of a hypothetical ”completely average” loan. The explanatory variables provide an adjustment to this baseline, either accelerating time to default by down shifting the survival curve or, conversely, increase the survival time and thus up shifting the survival curve. A simple example which differentiates between borrowers with different credit qualities is provided below to illustrate these effects.

5 Results

The results shown here illustrate the modeling approach and are simplified for clarity of exposition. We grouped FICO scores into three brackets:

- {660 - 724} → low
- {725 - 785} → mid
- {790 - 850} → high

The survival curves for each of the above FICO score brackets and the baseline survival curve are shown in Figures 2 and 3.

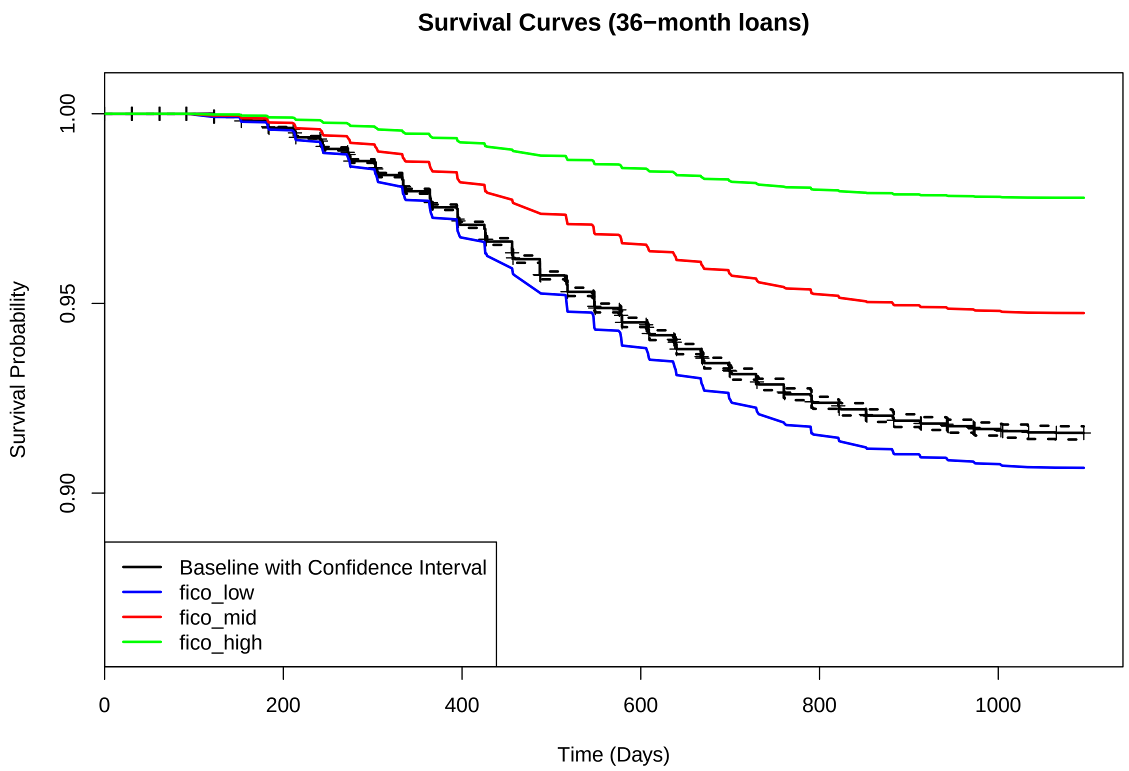

Figure 2: Survival Curves (36-month loans)

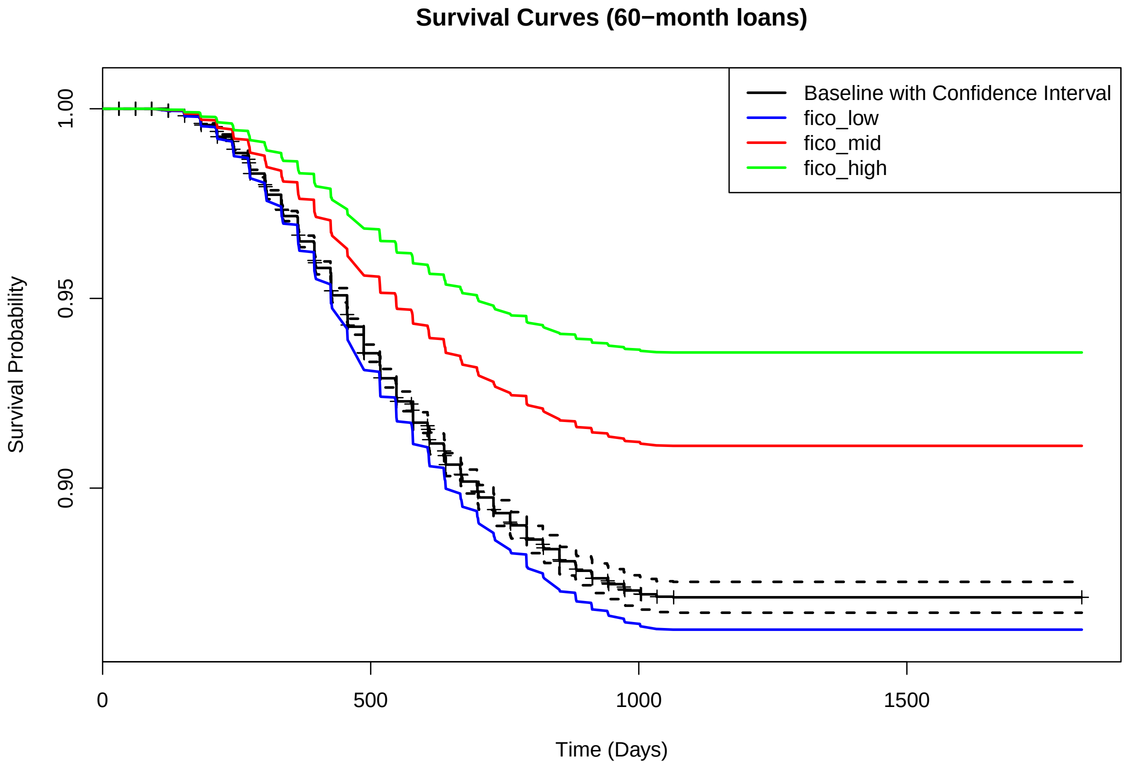

Figure 3: Survival Curves (60-month loans)

The tables below list the fitted regression coefficients together with the statistical significance (the p-values) for the 36 month and 60 month loans respectively. As expected, the coefficients monotonically increase with decreasing credit quality of the borrower. That is, higher coefficients lead to faster decaying survival curves and hence the probability of defaulting before the terminal date of the loan increases. A key factor in choosing to represent features in buckets is that all coefficients are significant at the 90% level.

	coef	exp(coef)	se(coef)	z	p
fico_low	1.48	4.38	0.15	10.09	0.00
fico_mid	0.88	2.41	0.15	5.89	0.00
fico_high	0.00	1.00	-	-	0.00

Table 1: Coefficients of the Cox proportional hazard model for 36 month loans.

We further note by comparing Tables 1 and 2 that the coefficients are lower for the 60 month loans, indicating within a credit quality bracket a high probability of default for 36 month loans in any given time horizon.

	coef	exp(coef)	se(coef)	z	p
fico_low	0.80	2.22	0.18	4.50	0.00
fico_mid	0.34	1.40	0.18	1.84	0.07
fico_high	0.00	1.00	-	-	0.00

Table 2: Coefficients of the Cox proportional hazard model for 60 month loans.

5.1 Which explanatory variables should I use?

There are a number of explanatory variables that can be added to the model, but the choice will depend on a number of factors. First, if the variable is categorical there should be sufficient observations of the variable in each category so that the coefficients corresponding to each of the dummy variables is statistically significant. Second, the variable should be meaningful to investors. For example, a Global Macro fund with an investment thesis that Silicon Valley tech professionals are a growth demographic is likely to be interested in geography in addition to year of employment as variables. Also, many of the fields have not been authenticated by the lending platforms and so they should be used with caution.

6 Conclusion

Institutional investors and investment managers seek to better characterize the credit risk of online consumer loans. This article described how to generate loan survival data and build a credit risk model that can be used for a number of applications including generating alpha, issuing protection and securitizing loans into bonds with the desired risk/reward profile. A simple example using FICO buckets was used to provide insight into the modeling approach. In the next blog article, we shall apply the credit risk model described here to estimate the DV01 leg of the loan and estimate the discounted fair value. We shall then proceed to include a pre-payment model.

References

[1] D. Cox and D. Oakes. *Analysis of Survival Data.* Monographs on statistics and applied probability. Chapman & Hall, 1996.

[2] M. Lin, N. Prabhala, and S. Viswanathan. Judging borrowers by the company they keep: Friendship networks and information asymmetry in online peer-to-peer lending. *Management Science*, 59(1):17–35, 2013.

[3] M. Tsai, S. Lin, C. Cheng, and Y. Lin. The consumer loan default predicting model - an application of dea-da and neural network. *Expert Syst. Appl.*, 36(9):11682–11690, Nov. 2009.