**Predicting Default in Peer-to-Peer Lending**

**Evan Generoli**

**Introduction/Background**

To be human is to be risk averse. Anecdotally, it is clear that humans are not big fans of risk and uncertainty. In fact, we seem to be hard wired to shy away from the unknown/uncertain. This seems especially true when it comes to our finances; investors in particular are averse to the unknown. While investors may sometimes be comfortable with risk, in general the risk must be proportional to the reward. However, the level of risk posed by a given investment is not always immediately clear. Given this situation, there is significant demand for estimating the level of risk for these types of investments.

One such investment type is Peer-to-Peer (P2P) lending investments. Peer-to-peer lending is a relatively new investment vehicle which allows average investors to invest in personal loans directly, rather than investing in a bank or savings and loan, who then issue the loans to borrowers. In P2P lending, potential borrowers apply for loans with the P2P lending company. If approved, these loans are listed on an online marketplace where investors fund the loans by buying small pieces of the loans (usually around 25$), called notes. In essence, the P2P lending company acts as a matchmaker, matching potential borrowers with investors via a loan marketplace. Potential borrowers post their loan request and personal information and investors decide which loans to invest in based on this information. The P2P lending company makes money by charging fixed fees to both borrowers and investors. This business model and the fact that these companies primarily exist online, leads to reduced overhead compared to a bank (the difference between P2P lending companies and banks is somewhat comparable to the difference between Uber/Lyft and taxi companies). These savings through reduced expenses can then be passed on to both borrowers and investors via more favorable terms (ie. better interest rates). In order to effectively choose which loans to invest in, investors may want to estimate how likely a particular borrower is to default based on the personal information they provided.

The largest and most popular company in the P2P lending business is Lending Club, which has been in operation since 2007. Lending Club has made available, on the data science website Kaggle, a complete dataset of all loans issued through their platform from their inception in June, 2007 through December, 2018. This dataset includes loan characteristics (ie. loan amount, interest rate, term length) as well as borrower characteristics (ie. annual income, length of employment, home ownership status). With this dataset, a model can be built to see which variables are predictive of loan default, as well as how risky of an investment a given loan is. That is, an individual loan’s likelihood of default can be predicted based on their associated set of covariates.

I therefore build a Bayesian binary logistic regression model to predict default in this dataset of Lending Club loans. I have identified a few questions of interest related to this analysis. First, does home ownership substantially reduce the probability of default? This is not meant in a causal sense, but rather a predictive sense, ie. home ownership status would likely not *cause* someone to default on their loan, but would be correlated with things that do and therefore would still be predictive of default. Second, how strong is annual income as a predictor of default? Clearly, annual income would have an inverse relationship with loan default. Third, is length of employment or age of oldest credit line a better predictor of default?

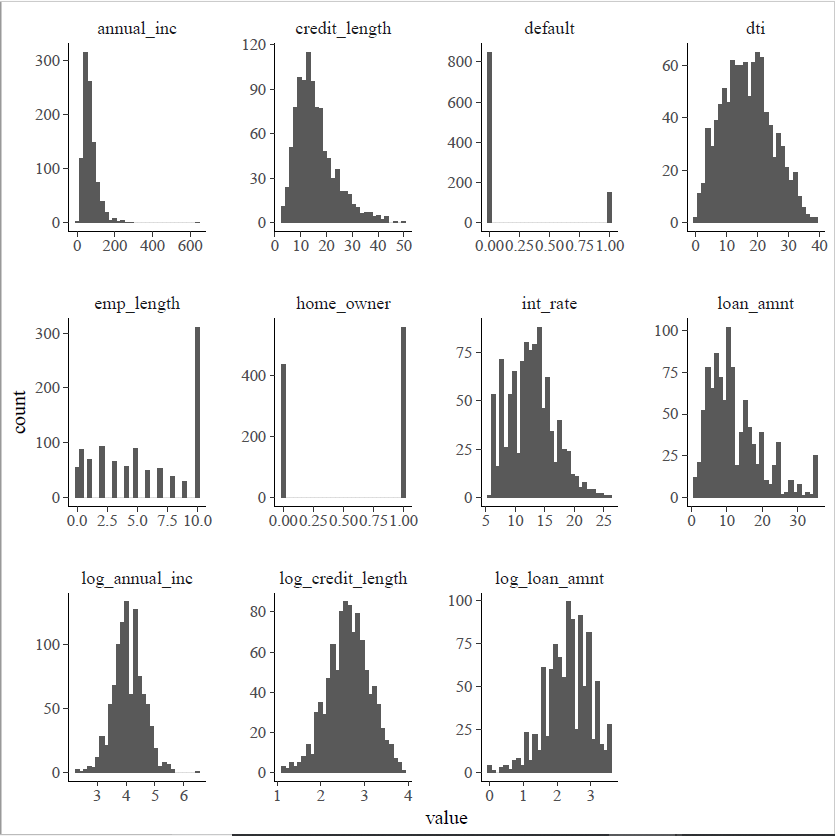
**Data**

As mentioned, the dataset for this analysis is a complete set of Lending Club loans and comes from Kaggle. The dataset as downloaded from Kaggle has 2,230,668 observations, but this includes many loans whose terms have not been completed yet ie. are still being paid or are currently late/in-default. This poses a problem, in that these loans’ history (or a part of it) is yet unwritten, ie. just because a loan is not currently in default doesn’t mean it never will be. In order to deal with this issue and simplify the analysis, I deal only with settled loans, ie. loan status is fully paid or written off (defaulted).

I therefore filter the dataset according to term completion date (issue date plus term length). I select only observations with a term completion date of December, 2017 or earlier. This date was chosen because it was the most recent term completion date for which there were no unsettled loans. That is, cutting the dataset at any more recent term completion date left unsettled loans, which would have to be dropped; this would be non-random data dropping, which would be a problem. After limiting the dataset as described, 358,894 observations remain. Finally, in order to make this analysis computationally feasible on a personal laptop, I take a random sample of 1000 observations from the truncated dataset of 358,894 observations.

Additionally, the initial dataset contained 145 variables of varying degrees of usefulness. To simplify the analysis, I have limited the dataset to only 8 variables of interest (including some which were created from existing variables): annual\_inc (annual income in thousands of USD), dti (debt-to-income ratio), home\_owner (an indicator for home ownership), credit\_length (age of oldest credit line), int\_rate (interest rate), loan\_amnt (loan amount in thousands of USD), emp\_length (length of employment in years), and default (an indicator of a loan status of “charged off”). Numerical and graphical summaries of the remaining variables are presented below. Note that logs of some of the variables have been included in the histograms.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table 1: Summary Statistics** | | | | | |
|  | Min | Median | Mean | Max | Std. Dev. |
| loan\_amnt | 1.00 | 10.00 | 12.26 | 35.00 | 7.49 |
| int\_rate | 5.42 | 12.86 | 12.77 | 25.80 | 3.84 |
| emp\_length | 0.00 | 5.00 | 5.52 | 10.00 | 3.75 |
| annual\_inc | 10.02 | 59.17 | 68.19 | 648.00 | 41.44 |
| dti | 0.26 | 16.21 | 16.74 | 39.09 | 7.95 |
| default | 0.00 | 0.00 | 0.15 | 1.00 | 0.36 |
| credit\_length | 3.17 | 14.17 | 15.76 | 49.83 | 7.68 |
| home\_owner | 0.00 | 1.00 | 0.56 | 1.00 | 0.50 |

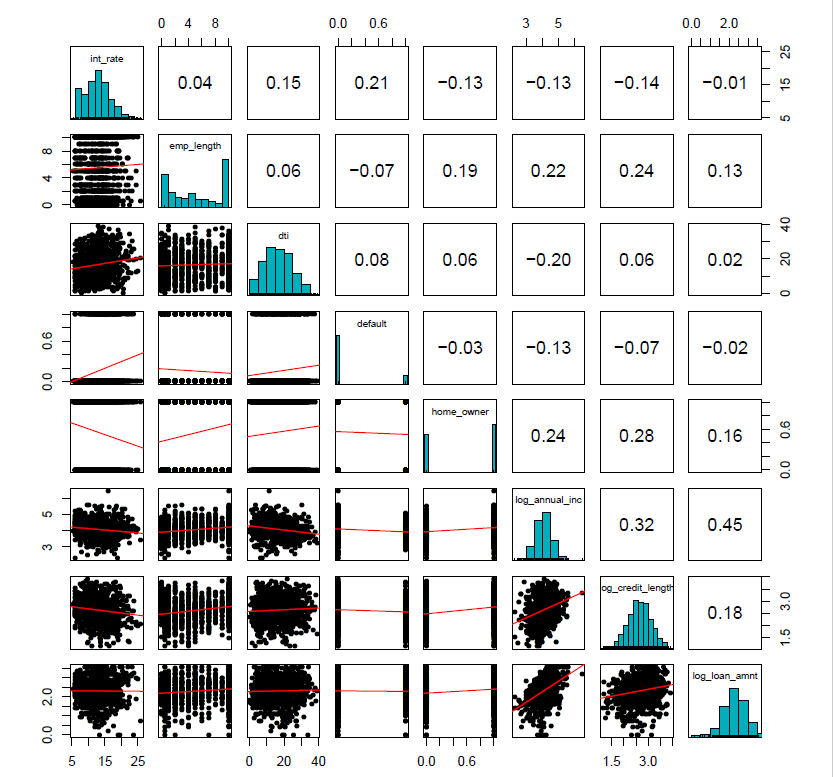
**Figure 1: Histograms of all variables** 

Note that the means of loan amount, annual income, and credit length are significantly greater than their respective medians, indicating positively skewed distributions. Additionally, the maximums of these variables are several standard deviations above the medians, further confirming that these distributions are skewed and could benefit from a transformation; this can indeed be seen in the histograms. Upon examination of these histograms, it is clear that annual income and credit length are benefitted by a log transformation, and loan amount appears somewhat helped as well. I therefore dropped the original (unlogged) version of all three of these variables and use the logs instead.

Additionally, note that the max of employment length is 10 years and that a large number of observations have this value. This is due to the fact that employment length was originally a categorical variable which included values of “10+ years” and “< 1 year.” These values were recoded as 10 and 0.5, respectively. While this is not an ideal solution, using the categorical employment length with 11 different categories would likely cause computational issues.

Finally, a matrix of correlation coefficients and pairwise scatterplots is displayed below. It is worth noting that there are quite a few correlations of substantial magnitude. In addition to the 0.45 and 0.32 correlations of log(annual income) with log(loan amount) and log(credit length), respectively, there are several correlations over 0.2 in magnitude

**Figure 2: Correlations & pairwise scatterplots of all variables**



**Methods**

In order to predict the probability of loan default, I fit several binary logistic regressions with a multivariate normal g-prior on the betas. Specifically, the data distribution is

where is the covariates for observation () and is the vector of parameters to be estimated. This is intentionally left in general form to accommodate fitting several models. The prior distribution on the betas is

where is an x diagonal matrix with elements and is calculated using the maximum likelihood (frequentist) estimates of the .

Given the binary nature of the response variable (default), it is clear that a Bernoulli GLM is the appropriate data distribution here. I chose the logit link function for this setting because it is extremely common and did not pose any issues (I tried the probit and complementary log-log link functions as well, but these were causing computational problems). I chose the multivariate normal g-prior because of the above-mentioned substantial correlation among predictor variables (ie. the matrix is not orthogonal, or even very close). Given this level of correlation present in the matrix, it is not reasonable to model the betas with independent priors. The mean of zero on the prior distribution is chosen to express prior ignorance and let the data speak for itself.

In order to determine which variables to include in the model, I execute a forward stepwise variable selection algorithm using the WAIC as a metric. The results of this selection algorithm are reported below. All models have been approximated using 3 chains of 10,000 iterations each (with the first half discarded as warm-up). The (Gelman-Rubin) statistics are all very close to 1, suggesting convergence of all the MCMC chains.

|  |  |
| --- | --- |
| **Table 2: Model Selection -- Forward Stepwise Selection** | |
| Variables in Model | WAIC |
| Interest Rate | 812.074 |
| Employment Length | 851.8296 |
| Debt-to-Income Ratio | 849.2093 |
| Home Owner | 855.4523 |
| Log(Annual Income) | 840.122 |
| Log(Credit Length) | 851.7088 |
| Log(Loan Amount) | 856.2157 |
| Interest Rate + Employment Length | 807.7218 |
| Interest Rate + Debt-to-Income Ratio | 810.838 |
| Interest Rate + Home Owner | 813.9563 |
| Interest Rate + Log(Annual Income) | 802.3916 |
| Interest Rate + Log(Credit Length) | 812.5074 |
| Interest Rate + Log(Loan Amount) | 813.5095 |
| **Interest Rate + Log(Annual Income) + Employment Length** | **801.8712** |
| Interest Rate + Log(Annual Income) + Debt-to-Income Ratio | 803.0758 |
| Interest Rate + Log(Annual Income) + Home Owner | 804.2344 |
| Interest Rate + Log(Annual Income) + Log(Credit Length) | 804.4707 |
| Interest Rate + Log(Annual Income) + Log(Loan Amount) | 803.5515 |
| Interest Rate + Log(Annual Income) + Employment Length + Debt-to-Income Ratio | 801.9076 |
| Interest Rate + Log(Annual Income) + Employment Length + Home Owner | 803.3039 |
| Interest Rate + Log(Annual Income) + Employment Length + Log(Credit Length) | 803.7277 |
| Interest Rate + Log(Annual Income) + Employment Length + Log(Loan Amount) | 802.6592 |

As can be seen in the model selection table, the model selected by the algorithm is the Interest Rate + Log(Annual Income) + Employment Length model. The Interest Rate + Log(Annual Income) + Employment Length + Debt-to-Income Ratio model is an extremely close second, but does not quite make it. Therefore, my final model is

Note the subscripts on the betas. This notation was used in order to match the stan output. To further validate MCMC convergence of my final model, I provide trace plots of the betas below.

**Figure 3: Trace plots of betas for final model**



To validate the chosen model, I perform posterior predictive checks with 1000 replicated datasets (). After extracting the MCMC chains for the beta posteriors, I randomly sample 1000 sets of betas (of 15,000 possible). From each set of betas, I create a replicated dataset by computing the log-odds (ie. ) for each observation and then transforming the log-odds into a probability () for each observation. In order for these posterior predictive checks to make much sense, each probability was then rounded to 0 or 1 in the standard way. A graphical summary of the replicated datasets is presented below in figure 4.

**Figure 4: Bar plot of observed data & medians**



It appears from the plot, that all of the replicated datasets are predicting close to zero defaults, ie. that nearly everyone will pay their loans. This does not appear reasonable relative to the observed data. In order to investigate this further, I define a test quantity equal to the number of defaults, that is, . I then apply this test quantity to my replicated datasets to generate and compare to the observed test statistic, . The results of this are shown below.

**Figures 5 & 6: Histogram of , alone & with red line of for comparison**

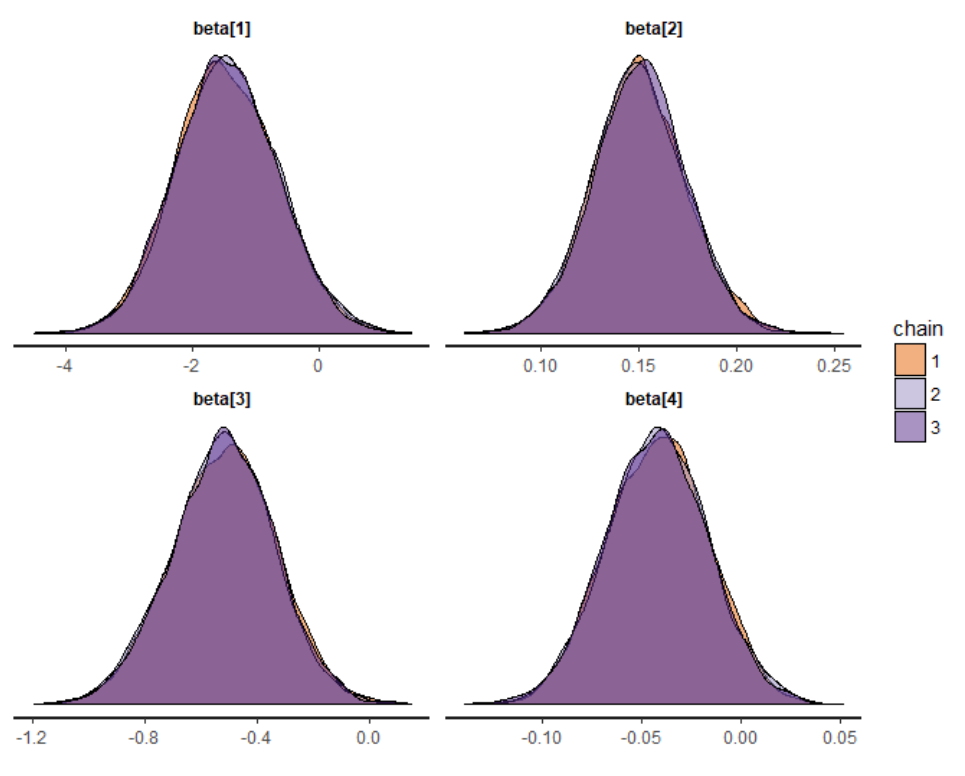
 

It can be seen that ranges from about zero to about 25, with the bulk of below 10. It is also quite obvious that all of is far below the 152 defaults in the observed data. This is clearly indicative of a model deficiency, as the model is massively underpredicting defaults (by an order of magnitude). Unfortunately, it is not clear what could be done to improve model fit without the addition of new variables.

**Results**

Posterior density plots of the betas are shown below. It can be seen from the density plots that (constant) is fairly negative, (interest rate) is unambiguously positive which makes sense, and (log annual income) & (employment length) are pretty clearly negative, which is also expected. Posterior summaries of the betas and exp(betas) are shown in table 3.

**Figure 7: Posterior density plots of betas**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Table 3: Posterior Summary of Betas & exp(Betas)** | | | | |
|  | Mean | Median | 2.5% | 97.5% |
| beta[1] | -1.451 | -1.469 | -2.996 | 0.158 |
| beta[2] | 0.151 | 0.15 | 0.104 | 0.198 |
| beta[3] | -0.522 | -0.517 | -0.892 | -0.162 |
| beta[4] | -0.042 | -0.042 | -0.092 | 0.008 |
| exp\_beta[1] | 0.325 | 0.23 | 0.05 | 1.171 |
| exp\_beta[2] | 1.163 | 1.162 | 1.109 | 1.219 |
| exp\_beta[3] | 0.604 | 0.596 | 0.41 | 0.851 |
| exp\_beta[4] | 0.959 | 0.959 | 0.912 | 1.008 |

Interpretation of the magnitude of these parameter estimates is best done with the means of the exponentiated betas. Note that interpretation of the constant is not particularly interesting, as it would be for a loan/borrower with an interest rate of zero, an annual income of $1000, and an employment length of zero. This would never happen in the real world, so I will ignore it. These results suggest that a 1 percentage point increase in interest rate is associated with a ~16% increase in the odds of default. This means that risk grows exponentially with interest rate, and that therefore the difference in risk between say a 7% and a 10% interest rate is actually quite large. It seems possible that the magnitude of this effect is due to interest rate being a derivative measure, ie. it is assigned based on estimate risk, and is therefore a proxy for other characteristics which predict default. These results additionally suggest that a 1 unit increase in Log(Annual Income) (ie. an increase by a factor of e≈2.7) is associated with a ~40% decrease in the odds of default. This means that a borrower earning $135,000 is 40% less likely to default than one earning $50,000. This estimate actually seems relatively small, ie. I would expect this difference in income to reduce risk more, but it seems likely that higher income earners also have higher loan amounts and therefore risk decreases only logarithmically with income. Finally, these results suggest that a 1 year increase in Employment Length is associated with a ~4% decreases the odds of default. This magnitude seems reasonable; the difference between 4 and 5 years of employment is not very stark, but the difference between 2 and 10 years is.

**Conclusion**

While the model fit, as evidenced by the posterior predictive checks, is not very good, this analysis can still be used to answer some interesting questions, including the research questions posed at the beginning of this paper. First, home ownership does not seem to be a particularly strong predictor of loan default. It was not strong enough to end up in the final model when performing stepwise selection, at least, which was somewhat surprising. Second, annual income is a quite strong predictor. An increase in income by a factor of e≈2.7 is predicted to decrease the probability of default by as much as 59%, or 40% on average. Finally, employment length seems to be a stronger predictor of loan default than age of oldest credit line. I found this result surprising, as age of oldest credit line would be highly related to overall credit score. However, it could be that interest rates are so correlated with credit scores that length of credit becomes less important after including interest rate.

Overall, this dataset/setting does not seem particularly well suited to Bayesian analysis. The computationally intensive nature of Bayesian statistics, ie. that every model fit must be simulated with multiple MCMC chains of thousands of iterations each, means that without access to a supercomputer, utilizing all that this dataset has to offer is near impossible. Recall that after limiting the dataset by term completion date to only settled loans, the dataset contained approximately 359,000 observations. While the frequentist binary logistic regression model had no trouble fitting on this many observations in a matter of seconds, the Bayesian MCMC chains began taking unreasonably long for anything over a few thousand observations, even for relatively simple models. In order to fit multiple models of varying degrees of complexity to perform model selection, etc., I had to limit the dataset to only 1000 observations (about 0.3% of the original sample). Even so, my stepwise selection algorithm took 30 minutes or more to run.

Moreover, I began running into issues of not having enough memory to allocate for assigning fitted models to R objects. Indeed, the MCMC chains table for my final model was ultimately too large to be extracted whole. I had to instead specify that only the betas were to be extracted, rather than include the transformed parameters (log-odds for each observation) and the generated quantities (log-likelihood for each observation). Furthermore, all of these computational issues arose in the context of a limited dataset of only 8 variables. Had I used all or even most of the 145 variables originally present in the dataset, and attempted to fit more complex models, these issues would have been compounded.

As discussed above, the model fit is not very good and needs to be improved to generate reasonable predictions. One possible area for improvement is the link function used. While I only used the logit link function in this setting for simplicity, there are no major issues posed by using a probit or complementary log-log link function (although the computation time would be longer). These models could be compared to logit-link model using the WAIC or LOOIC.

However, it is unlikely that merely changing the link function would have a drastic effect on model fit, particularly in light of the magnitude of the observed model discrepancy. It seems that utilizing more variables from the original 145 in the dataset would be required to significantly improve model fit. However, many of these variables were of questionable use and/or were categorical variables, which would significantly increase the number of parameters to be estimated, likely causing further computational issues. Nonetheless, utilizing 15 variables or so out of the 145 seems feasible and could likely improve model performance.

However, all of this seems like a lot to do to improve a severely deficient model. Even such an improved model would still suffer from serious computational issues compared to a frequentist model. Overall, given the large sample size and number of predictors present in this dataset, this analysis setting seems that it would be better served by a frequentist model.