Explaining Loan Default in Microfinance

Garrett Allen, Piper Hampsch, and Boxuan Li (4:45 Runtime)

Introduction

We aim to analyze loan data from the San Francisco based microfinance institution, Kiva. We rely on the Kiva database which provides a collection of information on a random sample of loans from 2005-2012 to explore the factors influencing loan default. Each loan issued to low-income individuals is connected to online lenders with a peer-to-peer investment model. Diving into this loan data will provide more context on the sustainability of Kiva's model and the future of its lending model.

We will discuss our data cleaning process, defining what is considered default in the dataset and which sociodemographic and geographic factors we expect to contribute to risk of default. We will explore the data visually to inform our modeling process, followed by a thorough exploration of models for both risk of default and time to default. We hope to use the exploratory data analysis and model output to make conclusions on which sociodemographic and geographic factors (from both the borrower and lender) contribute to risk of default, with which we hope to provide actionable insight to potential investors in Kiva.

Methodology

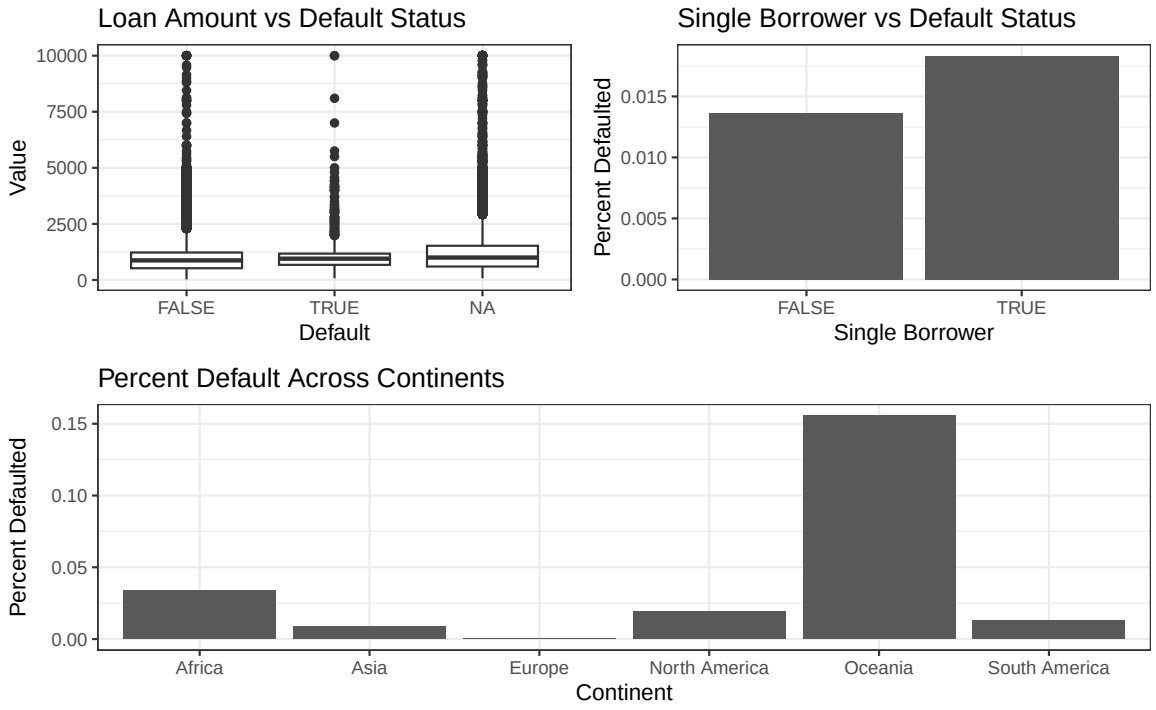
In order to understand what factors contribute to risk of default and time to default, we must first define default from the data available from Kiva. Though each loan has an assigned status, there are nuances between defaulted loans and paid loans that require further attention in defining default.

With our definition of default for this analysis, we hope to separate these labels into binary options: defaulted or not defaulted. After reading through Kiva's documentation on the status of loans, we will remove observations with statuses of expired, funded, fundraising, inactive, and refunded due to their smaller sample sizes and that these statuses do not have a clear binary option between defaulted and not defaulted. Loans in repayment require further separation,

and we will see how these loans align with their loan schedule in determining their risk of defaulting.

Exploratory data analysis

With the newly-defined variable for loan default, we can proceed with some exploratory data analysis on the response variable and potential predictors in the loans dataset. These visualizations will help inform the components of our model later on in the analysis.



Modeling Loan Default Status

It would be useful to Kiva's stakeholders to be able to predict whether a given loan will result in a default, allowing Kiva to advise borrowers on which loans are high risk. Kiva would become a better platform for investors who are concerned about losing their money. In order to explore this question, we will fit a logistic regression model that predicts the default status of a loan given various sociodemographic characteristics.

Each of our covariates has a good justification for adding it to the model. Our initial data analysis suggested that there may is a geographic correlation with probability of default, so we added continent to our model. We did not include country because it would add nearly

40 terms to the model, it ran the risk of overfitting, and continent worked similarly as well in predictive performance. The total loan amount seemed reasonable given that they directly relate to how a loan becomes paid off, and our EDA suggested differences in total loan amount between defaulted and non-defaulted loans. Our single borrower variable indicates whether or not the loan was given to a group of borrowers, or just one; this was added because it seems likely that a group of borrowers might be better at paying off a loan than just one. Type of geographic area also seemed reasonable to add, as our EDA showed that only defaulted loans are only in the country (as opposed to towns) making it a powerful predictor in our model for identifying true negatives.

There were some variables that, while they would have made the model perform better, we did not include because borrowers would not have access to them at the time of the loan (i.e. number of payments made on a loan, total payments made, etc.). We also found that the gender of the borrower was not a good separator of the data, even when considering how gender may interact with other variables in the model, like continent. Sector also was not a useful variable in the EDA, as the probability of default was similar within each sector. Interaction effects between the variables included were considered, but they did not help predictive performance and raised the AIC, so we omitted interaction terms from the model.

Modeling Loan Default

With all of this setup, we fit a logistic regression model using all of these covariates, with no interaction effects, to predict whether a loan is paid or in default. In repayment loans will not be addressed in this analysis, and they will be removed while fitting this model. In the next section, we will describe our methodology and assess model fit. In our results section, we will discuss how to interpret our coefficients, and in our discussion section, we will discuss implications of our model for which loans are likely to result in default.

Most of our coefficients have significant p-values, indicating that we can interpret their effects as significantly correlated with probability of a loan default. Backwards selection indicating that this model has the lowest AIC out of models with these covariates as the full model, and everything appears to be normal when looking at the coefficient output. VIF of all the covariates is low, with the worst VIF being 1.6. This is well below 5, so this is not particularly concerning.

Now that we have fit our model and no immediate issues seem apparent, we will perform model diagnostics on our model.

term	estimate	std.error	p.value	conf.low	conf.high
(Intercept)	0.0702	0.1144	0.0000	0.0560	0.0876
ContinentAsia	0.1933	0.0809	0.0000	0.1647	0.2261
ContinentEurope	0.0200	0.7098	0.0000	0.0033	0.0621
ContinentNorth America	0.7339	0.0869	0.0004	0.6175	0.8684
ContinentOceania	0.0000	1028.9124	0.9857	0.0000	0.0000
ContinentSouth America	0.4693	0.0847	0.0000	0.3966	0.5529
location.geo.leveltown	0.0000	177.5655	0.9169	0.0000	0.0000
single_borrowerTRUE	1.6325	0.0959	0.0000	1.3556	1.9742
loan_amount	0.9999	0.0000	0.0026	0.9998	1.0000

Model Diagnostics

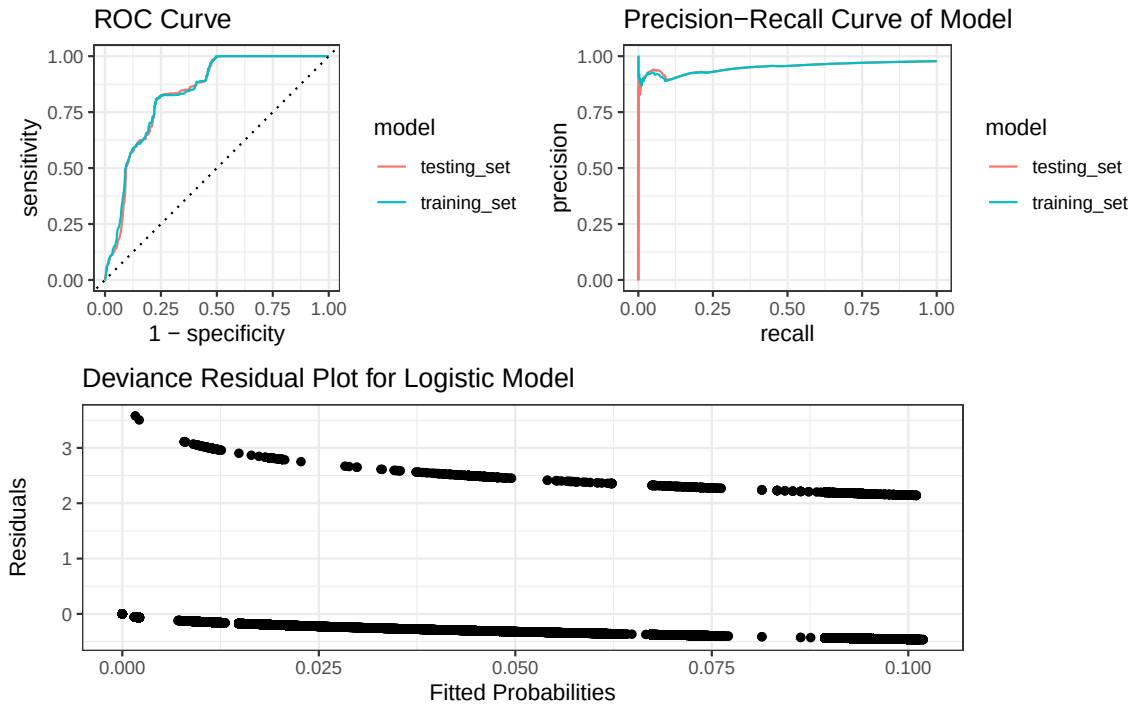
In this section, we will assess how well our model does predicting the default status of our loans. We broke up the data into a testing and training set with an 80-20 split, and we assessed how well the model performed using common logistic regression metrics like ROC curves, AUC, and PRC curves.

The AUC is around 90% for the training set and 89% for the testing set, which is quite high, and the deviance residual plot looks exactly like one would expect for a logistic regression model of this type. It has a clear pattern indicating the classes, with the residuals for the default class getting smaller as fitted probability grows, and residuals for the paid in full class getting smaller as fitted probability goes to 0. The ROC curves look very good for both the training and testing set, and same with the PR curves.

In short, the model appears to be performing moderately well on both the training and testing set. Sensitivity, recall, and precision are all moderately high, and our model appears to be performing better than random chance.

n	type
10244	true negative
3186	false positive
58	false negative
249	true positive

Above, we have have a table that shows the true positives, false positives, true negatives, and false negatives when the threshold for classifying a loan as defaulted is set at the best threshold (.03728), as determined by the ROC curve. We can see that our model is not performing very well, and that while it struggles with identifying true positives, the false negative rate isn't too bad, with around 18% of our truly positive defaulted loans in the test data set being classified as negative. Our model is doing significantly better than random chance, and thus we feel confident that we can interpret our model coefficients as being meaningful. We will give results on how to interpret the logistic regression in the results section of the paper.



Modeling survival time

We pre-determined several factors that may affect the time and probability to default and processed the data as follows:

1. We calculate the survival time as the total amount of time the borrower has been keeping up with the payments, calculated as the last payment processed date (roughly the time the borrower makes a payment) subtracted by the the first payment processed date.
2. We calculate the total number of days the loans get scheduled to be paid in full by the subtracting the first payment date from the last.
3. We calculate the percentage of the loan was paid by the borrower's last recorded payment (censored or failed). For borrowers who are paid, this value is 100%.
4. We calculate the percentage of time the borrower has been keeping up with the payment by dividing the borrower's total time of making the payment by the total amount of time for the loan to be paid. Note this value could be larger than one since some borrowers made their last payment after the due date of the loan.

For borrowers who have paid their loan, we know this borrower would not default through out the time span of the study, or the dataset. Hence, we set the last payment date of all borrowers who have paid in full to be 2012-03-02, one day after the last data was collected.

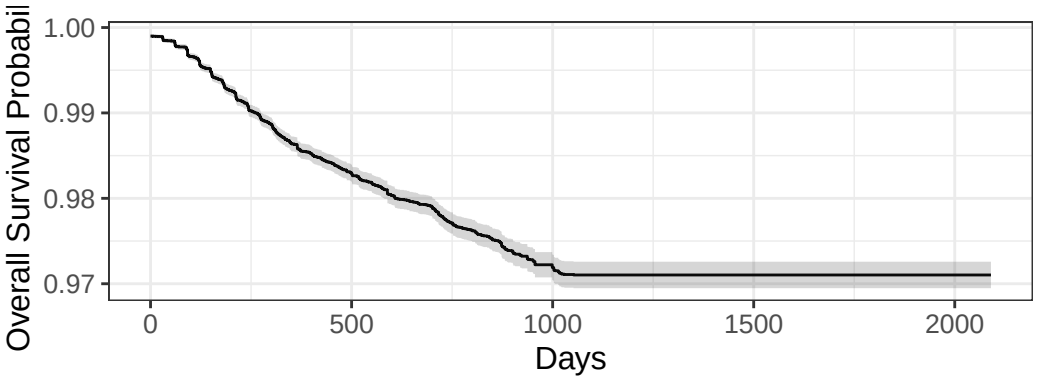
We fit a cox proportional hazard model using backward selection under a selection criteria of AIC. After elimination, the final model with the smallest AIC includes covariates of sector, geo-level location, nonrepayment loss and liability, borrower count, payment length schedule, payments count, continent, and time percent.

We first checked for linearity of the continuous/ordinal covariates. The continuous/ordinal covariates are payment length schedule, borrower count, and payments count. We plotted these covariates against martingale residuals of the null Cox proportional hazards model. From the plots, we saw that the linear relationship generally holds between the hazard and the log of payments count. We modified the model to include this. Since the linearity assumption does not hold for payment length schedule and borrower count, neither for the log of these covariates, we will drop them in the final model. We then checked the proportional hazard using the log-log curves. Due to page limitation, we could not show them in this report, but we created log-log visualizations for continent and sector.

Next, we check for proportional hazards assumption through survival curves. The lines are roughly parallel, so the assumption of proportional hazards generally appears to hold. Our final model is fitted with covariates of sector, geo-level location, nonrepayment loss and liability, borrower count, log-tranformed payments count, and continent. After eliminating the insignificant terms, the final model and its coefficient estimates are given by:

term	estimate
log(payments.count)	-0.9603
terms.loss_liability.nonpayment=partner	-8.3906
location.geo.level=town	-0.4323
sector=Clothing	-0.5826
sector=Construction	-0.4417
sector=Food	-0.6545
sector=Retail	-0.8433
sector=Transportation	-1.4586
sector=Wholesale	-1.4744

To help inform potential investors by estimating the probability of default, we use the Kaplan Meier estimate to calculate the survival curve, given by $S(t) = \Pi_{i:t_i \leq t}(1 - \frac{D_i}{N_i}) = \Pi_{i:t_i \leq t} \frac{S_i}{N_i}$. The survival curve is as follows. We can see most of the borrowers keep paying their loans. We present the 95% confidence interval and number of defaulted borrowers and those who are at risk as time passes. We also print out the risk of being default every half a year after the first payment.



Results

Results for Modeling Loan Default Status

For our continents variable, our baseline level was Africa, which every other continent had lower odds of defaulting than, as the sign of the coefficients on the log odds is negative. For example, the odds of defaulting in Asia compared to Africa with all other variables held constant decreased by a factor of .19, which is a rather large difference. Europe, South America, and Asia all had significant p-values at the .05 level, which suggests that the continent is correlated with the probability of default; lenders should broadly pay attention to the geographic location of where they are lending.

Being in a town as compared to the country resulted in default being much less likely; the odds of default decreased by a factor of of less than .0001, all else held constant. In fact, rather strangely, all of the defaults in our dataset here are in the country, making this by far the most important predictor of defaulting. This may suggest some sort of error with the dataset provided, as this seems unreasonable, but based on the data we have provided, type of geographic area was strongly correlated with probability of default. Thus, lending in towns appears to be significantly less risky than lending in the country.

Being a single borrower significantly raises the odds of default; specifically, being a single borrower raised the odds of default by 1.6, all else held constant. This suggests that being a single borrower is correlated with default, so lenders should pay attention to this variable when lending.

Each 1 dollar increase in the loan amount results in a slightly decreased odds of default, as the odds of default decreased by a factor of .9999. This effect does not appear to be as significant as the rest of the variables, but broadly suggests that more expensive loans may be correlated with lower default odds. This result should be interpreted cautiously; it may be that higher loan amounts are only funded when lenders think that the loans are more likely to not default, resulting in a selection bias in our dataset.

Results for Modeling Survival Analysis

For our model, our baseline for continent is Africa, nonrepayment loss and liability is lender, geo-level location is country, and sector is Agriculture.

For the Continent covariate, we can see that borrowers in most other continents have a smaller the hazard to default. For lenders in North America, the log hazard rate decrease by a factor of 0.5351, or have approximately 0.586 (exp(-0.5351)) times the hazard of in Africa, holding everything else constant. In South America, the hazard is 0.388 times that in Africa, and in Asia, 0.174 (exp(-1.7481) and in Europe, 0.015 (exp(-4.2252)) as the hazard in Africa, holding everything constant. For borrowers in Oceania, the p-value of the coefficient is not significant at 0.05 level. We also find a difference between sectors. For if the borrowed money is used for Clothing, Construction, Food, Transportation, and Wholesale, the corresponding hazard of default is 0.649, 0.558, 0.643, 0.520, and 0.430 of the hazard being in Agriculture, holding everything else constant. If the loss liability is the partner, the hazard is 0.382 of the hazard if the loss liability is the lender, and most notably, if the borrower lives in town, the hazard is only 0.0002 of that of the borrower lives in the countryside. If the log(payments.count) increase by one unit than another, the hazard will be 0.454 times of the hazard of the event than the other. For the rest of the coefficients, they are not significant at 0.05 level so we did not include them.

Discussion

Though we lay a foundation for analyzing loan default for Kiva, we identified some limitations to our work and possibilities for future work. In our logistic regression model, we intended to explore more recent data in the Kiva API so that we could pull in new information in light of the censored data. This would have allowed us to see if our logistic regression model performed well against newer loan data. Future work could carry out this intention and provide an in-depth model evaluation procedure with actual data from the Kiva API. We also considered exploring possible distributions for survival time, since we assumed it to be logistic in our analysis. Most of the coefficients were not significant for the survival model, so future work may look to explore the underlying distribution and consider more covariates in the model.

References

How does kiva work? Kiva. (n.d.). Retrieved January 20, 2023, from https://www.kiva.org/about/how

Bouquin, Daina. IS 608, (2016), GitHub repository, Retrieved January 22, 2023, from https://github.com/charlespwd/project-title