Improving Matching Between Interest Rates and Borrower Riskiness in Microfinance Loans Using Sampling Methods

Jack Willoughby*
Advisor: Dalene Stangl

June 10, 2015

Abstract

Increasing default rates coupled with high information asymmetry have the potential to discourage future lenders on Zidisha, an online microfinance platform, from offering money to borrowers at low interest rates. The goal of this analysis is to develop a solution to this emerging problem by improving the quality of information available to the lender about potential borrowers' riskiness prior to determining the interest rate at which he/she will offer to lend money. To do so, sampling methods will be used to build a posterior predictive distribution of loan outcomes as a function of the interest rate offered. The risk/reward preferences of prospective lenders will then be elicited and combined with predicted loan outcomes to propose the lender-specific optimal interest rate for a given loan. This is done in accordance with the goal to provide borrowers with a low cost of capital in the present while not leaving so many lenders dissatisfied with unprofitable loans that there will be a shortage of capital available for loan in the future.

*Jack graduated Duke University in May, 2015, with a B.S. in Economics with distinction and a B.S. in Statistics. He currently attends The Ohio State University where he is pursuing a Master's degree in Economics, after which he will work for McKinsey & Co. in New York City. He can be reached at jjwilloughby95@gmail.com.

1 Introduction

The purpose of this analysis is to improve alignment between interest rates offered by lenders and the true risk associated with microfinance loans on Zidisha.[1] Zidisha is an international nonprofit organization that serves as an online microfinance exchange on which borrowers from developing countries can apply for loans to be fulfilled by lenders worldwide. Borrowers then select which lenders will fulfill their loans based on the interest rates offered.

Lenders loan money to both help others and make money; the balance between these two goals varies by lender and drives the interest rate that a lender offers. The emerging problem with Zidisha is that increasing default rates, as illustrated by Figure 1, coupled with information asymmetry have the potential to discourage lenders from offering money to deserving borrowers at low interest rates. While lenders on Zidisha view their efforts as philanthropic, their long-term commitment to Zidisha may be conditional on feeling that their good will is being repaid by others: if people feel that their money is being stolen through fraud, they are less likely to offer money again in the future. Furthermore, information asymmetry makes it difficult for lenders to ascertain which loans are relatively riskier than others, and subsequently what interest rates to offer to satisfy their risk/reward preferences. Lenders can gather only limited information from prospective borrowers' descriptive loan requests and profiles because people with a low probability of repayment have an incentive to mimic the loan requests and profiles of people who successfully repay loans.

Previous studies have focused on determining interest rates to maximize the profit of the lender, but that goal does not translate to this analysis. Here, the goal is to minimize the cost of capital to borrowers, both now and in the long run. To do so, lenders must offer the lowest possible interest rate at which enough lenders will be satisfied with the outcome of their loan that there will continue to be capital available to borrowers in the future. While higher interest rates are not desirable, decreasing the cost of capital too severely in the short run, or creating a situation in which no lenders make money from their loans, has the potential to discourage future lending, which would increase future interest rates. To keep lenders satisfied, lenders willing to lose money, or those who value philanthropy relatively more in their utility

[1] Since the inception of this analysis, Zidisha has changed its platform to offer exclusively interest-free loans; the forthcoming model will apply only to the antiquated structure of Zidisha loan allocation.

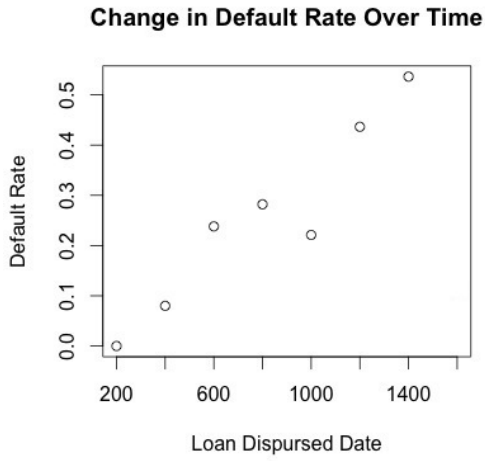

Figure 1: Change in default rate over time. Each point represents the average default rate for loans dispersed in the previous 200 days. Loan dispersed date is measured as days since the inception of Zidisha.

function, need to be matched with risky borrowers, and conversely lenders who desire return on investment need to be matched with borrowers with higher expected payback.

Using data from Zidisha, this analysis uses simulation to create a posterior predictive distribution of a loan's repayment rate. Interest rate is then incorporated with the predicted distribution of repayment rate to create a distribution of loan outcomes as a function of interest rate. Then, prospective lenders' risk/reward preferences are elicited, which are used to calculate an optimal range of interest rates for a lender and loan pair. This range will then be communicated to the lender. With more informative decision-making, lenders will be less surprised and more satisfied with their lending experience in the long run, which will keep interest rates low in the present and increase the long-term pool of capital available for deserving borrowers.

The paper will proceed as follows: Section 2 provides background information, Section 3 introduces the theory that underlies the analysis; Section 4 explains the data used; Section 5 explains the model and empirical methods; Section 6 discusses results and evaluates the model's predictive ability; and Section 7 concludes.

2 Background Information

Zidisha is an international nonprofit organization that serves as an online microfinance exchange on which borrowers from developing countries[2] can apply for loans to be fulfilled by lenders worldwide. To apply for a loan, borrowers create personal profile pages on which they can post a loan request with a description of how they intend to use the money upon its disbursal. Potential lenders can then scroll through the requested loans and choose on which, if any, they want to bid. Lenders bid on all or part of a loan with the interest rate at which they are willing to offer money, and borrowers ultimately select the lowest combination of interest rates to fill their loan, making the interest rate bidding process a form of reverse auction. After the borrower and lender or group of lenders have agreed on the interest rate, dollar value, and repayment period of the loan, the money is transferred instantaneously over the Zidisha platform without any intermediary required to handle or distribute the loan.

Zidisha supports loans with flat yearly interest rates that are allowed to vary from zero to 15%. In addition to the interest payments, borrowers must pay a fee to Zidisha of 5% of their loan's value per year the loan is held. If a lender fulfills only part of a loan, the share of the repayment he/she receives is equal to the share of the loan that he/she fulfilled. Zidisha does not, however, guarantee the repayment of loans, and instead makes clear that "lenders fund loans at their own risk and out of the goodness of their hearts, with the understanding that they may lose part or all of the funds lent."[3] While Zidisha encourages people to repay loans and says that they will use social pressures and legal remediation when necessary, there is little the company can do to force someone to repay his/her loan. In order to get a loan, however, people must not hold debt or have any outstanding loans, and borrowers cannot get a loan after defaulting a Zidisha loan. Borrowers' first loans can be up to $100 if they have been invited to join by a member currently in good standing, and up to $50 if not. The amount a borrower can request increases with his/her number of successfully repaid loans.

[2]Zidisha currently offers loans to residents of Benin, Burkina Faso, Ghana, Guinea, Indonesia, Kenya, Niger, Senegal and Zambia.

[3]www.zidisha.org/faq

3 Theory

Unlike most lending problems in which lenders select the profit maximizing basket of loans and interest rates from a pool of people of varying riskiness, here the goal is not to maximize the lender's net payoff. Instead, the goal is to minimize the interest rate that borrowers must pay such that enough lenders are sufficiently satisfied with their loans to continue to offer money in the future, and thus ensure that future borrowers will receive cheap capital as well. To determine this optimal interest rate, the quantified predicted riskiness of a loan must be combined with elicited preferences of potential loaners.

In order to quantify the true riskiness of a loan, the data analyst must determine the full predictive distribution of loan outcomes, which incorporates the size of a loan with the probability of default and the distribution of repayment conditional on default; all default is not valued equally by loaners, and therefore should not be considered equal by the data analyst. The predictive distribution of loan outcomes must then be combined with the lender's utility function to either determine the lender's optimal range of interest rates or inform the lender that there is no feasible interest rate that will make a loan satisfactory to the lender. Diminishing marginal utility of money implies that there is a non-linear relationship between the value of money and the value of money to people, which is why \$100 is a lot more valuable to a person with only \$1,000 in his/her bank account than it would be to that person if he/she were a billionaire. Additionally, varying levels of risk aversion predict that lenders will value loans with equal expected returns differently if they are associated with different risk levels. Consider the following decisions:

Decision 1: Bet $\$B$ on a fair coin flip?

Decision 2a: Take $\$B$ guaranteed or gamble and receive 0 with probability = 0.9, $10 \times \$B$ with probability = 0.1?

Decision 2b: Take $\$(B + C)$ guaranteed or gamble and receive C with probability = 0.9, $C + 10 \times \$B$ with probability = 0.1?

Rates at which people would elect to participate in one or all of the gambles would vary as B and C vary between -\$10,000, -\$1, \$0.10, \$1 \$10, and \$10,000, even though for given values of B and C, the expected profit of the

decision maker is the same regardless of the decision made. This non-linear relationship between the value of money and the value of money to people requires the calculation of the entire distribution of loss conditional on default, not just its expected value or the probability of default.

Next, the repayment of a loan is likely correlated with its interest rate in two ways. First, there might exist signals available to the lender but not to the data analyst, like the quality of the grammar used in the loan application or the subjective plausibility of the idea to be funded by the loan, that correlate with the probability of default and influence the interest rate offered. For example, a poorly written loan request will have a higher probability of default than the average loan and therefore likely be fulfilled with a relatively high interest rate. Observation would indicate that the loan was defaulted and was offered a high interest rate, but the conclusion that the high interest rate caused default would likely be misplacing causation; instead, it was the borrower's quality that caused both default and the high interest rate. Therefore, including interest rate in the model exposes it to severe confounding stemming from unquantifiable signals perceivable to the lender but not the data analyst.

Second, the interest rate may affect the loan outcome if it alters the borrowers ability or desire to repay a loan. For example, a borrower who initially satisfies high interest payments may reach a point at which he/she does not have enough money left to complete the project, and therefore a high interest rate partially causes default. Similarly, it is conceivable that someone offered a very low interest rate will perceive that the lender does not value money very much, and will therefore consider the loan philanthropy and feel less morally compelled to repay the loan. The effect of interest rate on a loan is theoretically difficult to determine, and empirically impossible to determine because of the confounding with interest rate selection. To determine the effect of interest rates on loan outcomes, Zidisha would need a subset of the loans to have been randomly assigned interest rates, but this unfortunately does not exist. To avoid the confounding stemming from subjective selection of interest rates, which is likely the biggest influence on the correlation between interest and default rates, the interest rate of a loan will be considered independent to its eventual repayment. This implicitly assumes that a given borrower will repay his/her loan at the same repayment rate regardless of the interest rate offered, which is a limitation of the model.

Therefore, the posterior repayment distribution will be calculated independent of interest rate, and then interest rate will be incorporated to find

a posterior predictive distribution of loan outcomes as a function of interest rate such that:

$$\text{Loan outcome} = \text{repayment rate} \times (1 + (r \times \frac{\text{loan length}}{12 \text{ months}})) \times \text{principal} \quad (1)$$

Different probability and repayment outcome combinations will then be elicited from individual borrowers and will be used to determine a range of interest rates that will, on expectation, make the borrower indifferent between fulfilling and passing on the loan. The resulting recommended interest rate will be the minimum rate such that enough lenders are sufficiently satisfied that they will continue to lend money in the future. If there is no interest rate below the borrower's stated maximum accepted rate that will satisfy the lender, the lender will be informed that he/she can only bid on this loan if he/she is willing to adjust his/her risk/reward preferences.

4 Data

The data used in this analysis were recorded internally by Zidisha and came in two parts: information on individual loans and background information on each borrower. First, the data was aggregated at the loan level, so that each observation is a unique loan with an associated borrower. Some borrowers have taken more than one loan, so there exist some loans that share the same background borrower information. This aggregation yielded 7184 loans. Next, all loans that were still outstanding at the time of data collection were removed from analysis. These loans have not yet been completed, and could result in either default or repayment. This censoring is a potential source of bias if the reason loans were still outstanding was correlated with their outcome, which is a limitation of the analysis. After their removal, the data were comprised of 3185 total loans with 2625 unique borrowers. The next step was to code default, which is defined as either failing to repay a loan within 6 months of its repayment date or not making or rescheduling any payments for a period of 6 consecutive months. Every loan is associated with a date, and since people are not allowed to receive another loan after defaulting, the last loan made by any person who has defaulted must be the defaulted loan. By the same logic, all previous loans for that individual must have been fully repaid. The completed loans associated with borrowers who are listed as never having defaulted were coded as fully repaid. Using this

Loan Number	N	Default Rate
1	2625	.3825
2	479	.1879
3-5	81	.1728

Table 1: Sample size and default rate by borrower loan number.

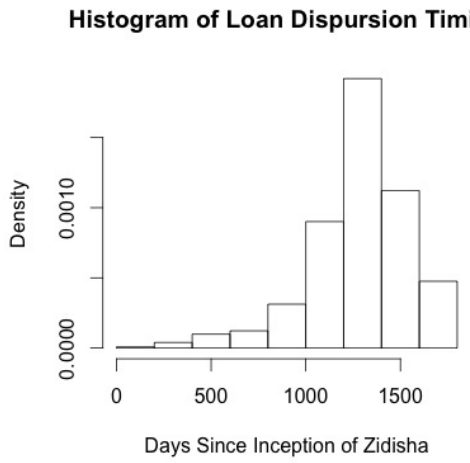

Figure 2: Histogram of the Dispersion date of completed loans, measured in time since the inception of Zidisha. The recent tapering is due to the fact that many of the most recent loans are currently outstanding and therefore do not enter the data.

methodology, default was coded for all loans. The aggregate default rate is 34.79%, which is decomposed by individuals' loan number in Table 1. In analysis, loan number categories 1 and 2 correspond to a borrowers first and second loan, respectively, while category 3 includes borrowers' third to fifth loans. Additionally, as evidenced by Figure 1, the default rate has increased over time. Finally, participation in Zidisha has also increased over time, as demonstrated by Figure 2.

Along with default rate, this analysis is interested in repayment rates. Figure 3 illustrates the distribution of repayment rates conditional on default and nonzero repayment. This histogram does not include the subset of defaulters who repaid none of their loan, which occurred in 35.2% of all defaults. Ultimately, this analysis aims to recommend interest rates, a his-

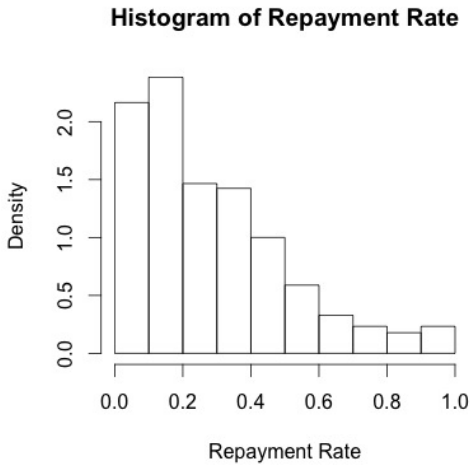

Figure 3: Histogram of repayment rate conditional on default and that repayment rate > 0.

togram of which is shown in Figure 4. In predicting the effect of interest rates, interest rates will likely be endogenous since there might exist signals available to the lender but not to the data analyst, like the plausibility of the idea cited in the loan application, that correlate with both probability of default and interest rate offers. The relationship between interest rates and default rate is shown in Figure 5.

One downside of the data is that there are some variables that cannot be implemented into the model because they are missing data for the vast majority of the loan observations. For example, only 6.53% of loans correspond to borrowers who were invited to Zidisha by another member of Zidisha. Conditional on a borrower's being invited to Zidisha, there exists data on the borrower's inviter, like his/her own loan repayment rate. This data is only available for the 6.53% of the loans for which an inviter is present, and therefore 93.47% of loans have missing data for inviter-specific variables. Furthermore, some of these variables are perfectly correlated with other covariates and whether or not a borrower defaults. Therefore, the model will include only a binary variable that indicates whether a prospective borrower was invited to Zidisha, and another to indicate whether he/she invited others to the microfinance exchange. Additionally, 77.6% of all loans are from borrowers who reside in Kenya, while the remaining 22.4% are from 8 different

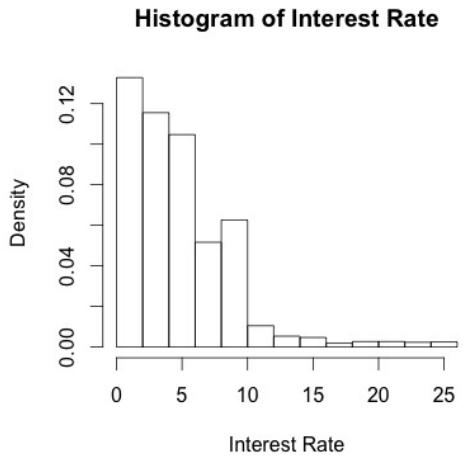

Figure 4: Histogram of interest rate of loans. For loans funded by more than one lender, this is an average rate weighted by the share of the loan funded at each interest rate. Note that interest rates could previously vary from zero to 25%, but now are limited to the range from zero to 15%

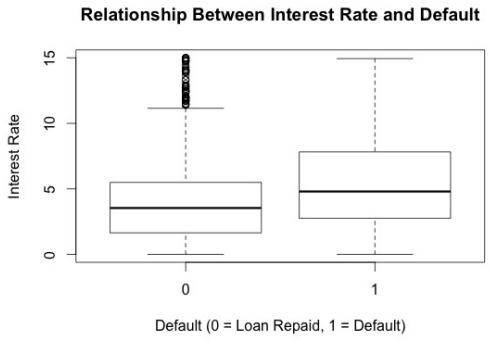

Figure 5: Boxplot of interest rates grouped by loan default status.

countries. These remaining countries do not each have enough observations to reliably indicate between-country variation and, in some cases, are perfectly correlated with other variables, so country of origin was simplified to a binary variable that indicates whether or not a borrower is from Kenya.

Finally, the data will be randomly split into a training and testing set. The training set will be used to fit the model, or to predict the influence of covariates on default and repayment conditional on default, and the testing set will be used to test the model. The interest rates and outcomes of the testing set will be thrown out and their loan outcomes will be predicted using the methodology explained in section 5, and then true and predicted outcomes will subsequently be compared to assess the predictive power of the model.

5 Model and Empirical Methods

5.1 Determining covariate influence with training data

Each testing set loan's posterior predictive probability distribution of repayment will be calculated using sampling methods and the statistical software R. The first step in the model is to use the training set to calculate the effect of different characteristics of a loan and borrower on the loan outcome. This will be done with four regressions, each of which will be used to generate a vector of β coefficients. The first of these regressions will fit a binary dependent variable that indicates whether a person has defaulted the loan:

$$Y_{1,i} = \begin{cases} 1 & \text{if loan is defaulted} \\ 0 & \text{otherwise} \end{cases}$$

The regression will utilize a logit transformation and predict the default outcome using available covariates known before the loan is disbursed, like country of origin, number of characters used in writing the loan application, whether the person has listed references, the loan number of the person, the time since the person has joined Zidisha, and the loan size. The model fitted will be:

$$\begin{aligned} &\Pr(Y_{1,i} \mid X_i) = p_i \text{ where} \\ &\log(p_i/(1-p_i)) = X_i^T \mu_1 \end{aligned} \tag{2}$$

This will be fitted with all training data to find μ_1, or the vector of point estimates of the impact of covariates X on whether someone defaults, and

Σ_1, the corresponding correlation matrix between logistic regression parameters. The resulting MLE values contained in the vector μ_1 are estimated with varying levels of certainty, which should be reflected in their use in prediction of probabilities of default for testing data. To reflect this uncertainty, a vector $\beta_{1,k}$ will be drawn for each simulation for the testing loans, such that all testing loans will share the same $\beta_{1,k}$ for each simulation k. The k different β_1 vectors will be drawn from a multivariate normal distribution with mean μ_1 and covariance matrix equal to the correlation matrix between the regression parameters of the logistic regression, which will be extracted from the regression results from Equation 2.

The second regression will fit a binary dependent variable that indicates that the repayment rate conditional on default is precisely zero:

$$Y_{2,i} = \begin{cases} 1 & \text{if loan is defaulted with zero repayment} \\ 0 & \text{if loan is defaulted with nonzero repayment} \end{cases}$$

This draws a distinction between loans that are defaulted but partially repaid, and those that are defaulted with zero repayment. Similar to the regression used to predict default, the regression used to predict when repayment is zero conditional on default will utilize a logit transformation. The model fitted will be:

$$\begin{aligned} &\Pr(Y_{2,i} \mid X_i) = q_i \text{ where} \\ &\log(q_i/(1-q_i)) = X_i^T \mu_2 \end{aligned} \tag{3}$$

This model will be fitted with the subset of the training set loans that were defaulted to find μ_2, or the vector of point estimates of the impact of covariates X on whether someone repays none of their loan conditional on default. Similar to above, the vector β_2 that will be used to calculate probabilities of zero default conditional on default for testing set loans will be drawn from a multivariate normal distribution with mean vector equal to μ_2 and covariance matrix equal to the correlation matrix between the regression parameters in the logistic regression in Equation 3.

The subset of training data loans that are not fully but at least partially repaid will be used in two subsequent regressions to predict repayment conditional on both default and partial repayment. The first will use the outcome variable of repayment conditional on default but partial repayment, R_i. Since $0 < R_i < 1$, to ensure normal residuals when using OLS regression, prior to being modeled R_i will be transformed to vary between negative and positive

infinity:

$$Z_i = \log(R_i/(1 - R_i))$$

A model to predict the influence of covariates on Z_i will subsequently be fit using JAGS in R. While typical regression models assume that the error term is uncorrelated with the predictors, this model will assume normal residual error that depends on the covariate values. This will be done because it is conceivable that the riskiness of a loan, or its divergence from its expected repayment rate, varies in correlation with the covariates, and because of the nonlinear relationship between the value of money and the value of money to people, variations in deviation from the mean repayment value are important to prospective lenders. The resulting model used to predict the influence of covariates on Z_i is as follows:

$$\begin{aligned} &Z_i = X_i^T \beta_3 + \epsilon_i \text{ where} \\ &\epsilon_i \sim \mathrm{N}(0, \sigma_i^2) \text{ with} \\ &\log(\sigma_i^2) = X_i^T \beta_4 + \delta \end{aligned} \tag{4}$$

Where $\delta \sim \mathrm{N}(0, \gamma^2)$. This regression will determine β_3, or the impact of covariates X on an individual's repayment rate conditional on default but partial repayment, and β_4, which quantifies the effect of covariates on the variability of the loan.

5.2 Calculating loan specific values for simulation

For each simulation, k, draw $\beta_{1,k}$ from a multivariate normal distribution with mean equal to the vector of MLE point estimates from the logistic regression predicting default, and the covariance equal to the correlation matrix between the regression parameters in that logistic regression:

$$\beta_{1,k} = \mathcal{N}(\mu_1,\ \Sigma_1) \tag{5}$$

The result will be a matrix of k different β_1 vectors that will be used to calculate $p_{j,k}$ for simulation k of each j testing set loan.

Similarly, for each simulation k, draw $\beta_{2,k}$ from a multivariate normal distribution with mean equal to the vector of MLE point estimates from the logistic regression predicting zero repayment conditional on default, and the covariance equal to the correlation matrix between the regression parameters in that logistic regression:

$$\beta_{2,k} = \mathcal{N}(\mu_2,\ \Sigma_2) \tag{6}$$

The result will be a matrix of k different β_2 vectors that will be used to calculate $q_{j,k}$ for simulation k of each j testing set loan.

For each testing set loan j, combine known covariates X_j with the calculated β_3 values to calculate loan j's mean predicted loan specific repayment rate conditional on partial repayment, V_j:

$$V_j = e^{X_j^T \beta_3}/(1 + e^{X_j^T \beta_3}) \tag{7}$$

Finally, combine covariate values and β_4 to calculate σ_j, which is the loan specific mean variation associated with repayment rate conditional on default and partial repayment:

$$\sigma_j^2 = e^{X_j^T \beta_4} \tag{8}$$

These point estimates will be used in the sampling model that will generate a posterior predictive distribution of repayment rate for each loan.

5.3 Simulating a posterior

After using available training set data to calculate μ_1, Σ_1, μ_2, Σ_2, β_3, and β_4, combining μ and Σ values to draw $\beta_{1,k}$ and $\beta_{2,k}$, and combining β_3 and β_4 values with known covariates for a testing set loan j to determine V_j and σ_j, simulation will be relied upon to create posterior predictive distributions of loan repayment rates for the testing set. The simulation process is enumerated below and illustrated in Figure 6.

1. Combine the drawn $\beta_{1,k}$ with testing set loan j's known covariates X_j to predict its probability of default for simulation k, or $p_{j,k}$, such that $p_{j,k} = e^{X_j^T \beta_{1,k}}/(1 + e^{X_j^T \beta_{1,k}})$.

2. Combine the drawn $\beta_{2,k}$ with testing set loan j's known covariates X_j to predict its probability of zero repayment conditional on default for simulation k, or $q_{j,k}$, such that $q_{j,k} = e^{X_j^T \beta_{2,k}}/(1 + e^{X_j^T \beta_{2,k}})$.

3. Draw $D_{1,j,k} = \text{Bernoulli}(p_{j,k})$ to determine if simulation k for testing set loan j was defaulted. If $D_{1,j,k} = 0$, the simulated loan is repaid in full. If $D_{1,j,k} = 1$, the simulated loan is defaulted.

4. If simulated loan was repaid ($D_{1,j,k} = 0$), record a simulation of repayment $R_{j,k} = 1$ into outcome space and return to step 1.

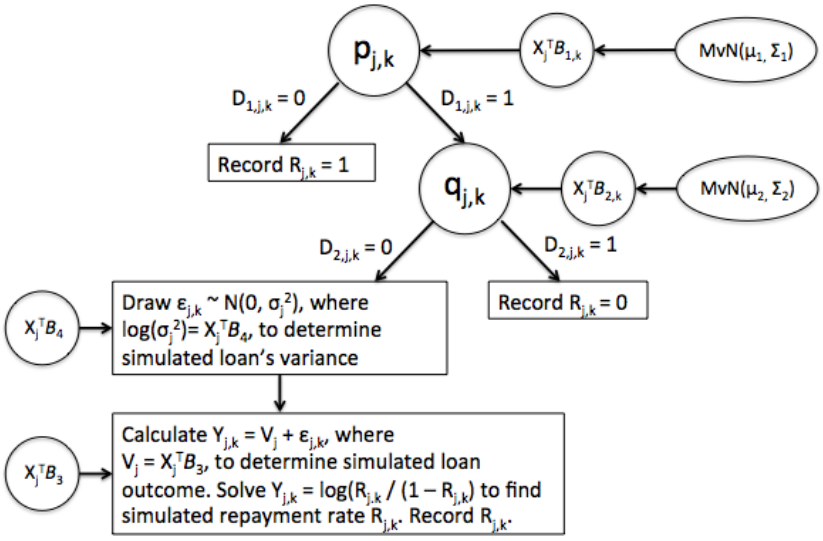

Figure 6: Simulation model

5. If simulated loan was defaulted ($D_{1,j,k} = 1$), draw $D_{2,j,k} = \text{Bernoulli}(q_{j,k})$ to determine if the simulated defaulted loan resulted in zero repayment. If $D_{2,j,k} = 0$, the simulated defaulted loan is partially repaid. If $D_{2,j,k} = 1$, the simulated loan is defaulted with zero repayment.

6. If simulated loan was defaulted with zero repayment ($D_{2,j,k} = 1$), record a simulation of repayment $R_{j,k} = 0$ into outcome space and return to step 1.

7. If simulated loan was defaulted with a nonzero repayment rate ($D_{2,j,k} = 0$), use testing set loan j's calculated mean loan variance, σ_j, to draw $\epsilon_{j,k} \sim \text{N}(0, \sigma_j)$ to determine the simulated loan's variance.

8. Use testing set loan j's calculated mean repayment rate conditional on partial repayment, V_j, and its drawn simulation specific variance, $\epsilon_{j,k}$, to calculate the loan outcome $Y_{j,k} = V_j + \epsilon_{j,k}$.

9. Solve the equation $Y_{j,k} = \log(R_{j,k}/(1 - R_{j,k}))$ for $R_{j,k}$ and record $R_{j,k}$ into outcome space.

10. Repeat for k = 10,000 simulations for each testing set loan j.

The simulation process outlined above will yield a posterior predictive distribution of repayment rate that is bounded $0 \leq R_{j,k} \leq 1$. The final step is to incorporate interest rate to convert the posterior predictive distribution of repayment rate to a posterior distribution of loan outcomes. To do so, each simulated repayment rate in the distribution will be multiplied by the loan size as done in Equation 1:

$$\text{Loan outcome}_{j,k} = R_{j,k} \times (1 + (r_j \times \frac{\text{loan length}_j}{\text{12 months}})) \times \text{principal}_j$$

This scaling will yield an array of loan outcomes as a function of interest rate; all inputs on the right side of the equation are known excluding interest rate.

5.4 Eliciting lender-specific risk/reward preferences

Once the posterior predicted distribution of loan outcomes is created for a given loan, risk/reward preferences of a specific lender must be elicited by asking the lender to identify values of θ that satisfy statements like:

- On average, I want my loan to yield a profit of $\$\theta$.
- I want a $\theta\%$ probability of my loan returning positive profit.
- I am willing to lose more than $\$\theta$ with only a 5% probability.

Each of these probability and loan outcome combinations from an individual borrower will correspond to a unique interest rate that can be used to create a range of interest rates for Zidisha to quote to the borrower. Zidisha does not necessarily want to recommend the interest rate that will actually make the lender as well off as he/she desires. Lower interest rates would make current borrowers better off but leave more lenders dissatisfied with their loans, which would decrease the subset of current loaners who make future loans. Conversely, higher interest rates would make current borrowers worse off but make more lenders satisfied with their loans, which would increase the subset of current loaners who make future loans. The share of lenders whom Zidisha wants to make satisfied with their lending experience and continue to offer loans in the future should reflect the projected growth of the pool of lenders vs. borrowers. If Zidisha believes, for example, that in the future there will be an influx of lenders but no change in the number of borrowers,

it would advocate low interest rates today because losing current lenders to dissatisfaction would not result in increased future interest rates since they would be replaced by future lenders. A converse argument can also be made.

Forecasting future changes in the pool of capital supplied by lenders and demanded by borrowers is beyond the scope of this analysis; instead, assume that lifetime interest rates are minimized by communicating to lenders the indifference-promoting interest rate associated with his/her stated preferences. Therefore, Zidisha will recommend a range of interest rates that will, on expectation, make the borrower indifferent between fulfilling and passing on the loan. If there are multiple interest rates that satisfy a lender's risk/reward preferences, Zidisha will quote the lowest of the possible rates to ensure that all else equal for the lender, the cost of capital is as low as possible to the borrower. If there is no interest rate below the borrower's stated maximum accepted rate that will satisfy the lender, Zidisha will communicate to the borrower that he/she can only bid on this loan if he/she is willing to adjust his/her risk/reward preferences. Lenders who desire return on their investments will pass on risky loans and seek out loans with higher posterior predicted distributions of repayment, leaving the higher risk loans for lenders who value philanthropy more than return on investment.

6 Results

The model outlined above was fitted with the training data to generate a collection of β vectors, which were then inputted into the sampling model to produce posterior distributions of loan repayment rate for each testing set loan. The model is designed to create what can be thought of as an "a priori" distribution of loan outcomes, or the true distribution of loan outcomes that exists prior to the disbursement of a given loan. Unfortunately, this a priori distribution of loan outcomes is forever unknown; each loan results in only one outcome. The challenge in assessing the quality of model fit is therefore to determine how accurately the sampled posterior distribution matches the true a priori distribution using only one outcome observation for each loan.

Actual loan repayment rates of the testing set are reintroduced to test the quality of fit of the model. The first means of evaluating the success of the model is to look at the confusion matrices for the logistic regressions predicting default and zero repayment conditional on default, shown in tables 2 & 3 respectively. These matrices display how successfully each model

predicts its outcome variable for the testing data, as well as the rates of false positives. For the logistic regression predicting default, the cutoff point for prediction was set such that about 35% of loans would be predicted to be defaulted, which matches the share of training set loans that were defaulted. This resulted in a cutoff of $\Pr(Default) = 49.5\%$, or that all testing loans with a predicted probability of default greater than 49.5% were predicted to have been defaulted. While this cutoff led to roughly the correct number of predictions of default, the model was prone to confusion and able to predict default vs. repayment with only moderate success. Default only occurred in 63.6% of testing loans that were predicted to default, while repayment occurred in 80.6% of loans that were predicted to be repaid.

The regression predicting zero repayment conditional on default performed slightly worse than the regression predicting just default. The probability of zero repayment conditional on default for training loans is roughly 37%, and the cutoff point for prediction attempted to match that rate. This is not possible to do directly, however, since the number of defaults in the testing set is not known prior to the determination of the cutoff value. To circumvent this challenge, the 35% of testing loans with the highest probability of default are considered the "defaulted" loans, and the cutoff was set so that roughly 37% of those loans are predicted to default with zero repayment. This resulted in a cutoff of $\Pr(R = 0 \mid Default) = 55.4\%$, or that all testing loans with a predicted probability of zero repayment conditional on default greater than 55.4% were predicted to have repaid none of their loan. This method turned out to overpredict zero repayment conditional on default by nearly 18%. Moreover, only 48.4% of the testing loans predicted to have zero repayment actually did have zero repayment, while 79.3% of loans predicted to be partially repaid were, in fact, partially repaid. In part because the cutoff for zero repayment conditional on default was set too low, the model does not seem to predict zero repayment very well.

The model fit was also tested to determine how empirically well calibrated it is. If the model is empirically well calibrated, then a loan's true outcome will be less than the α percentile of the loan's posterior predictive distribution for roughly $\alpha\%$ of the testing loans. For example, about 25% of testing loans should have a true repayment that is less than the 25th percentile of their posterior distribution for the model to be empirically well calibrated. Here, the extent to which the model is empirically well calibrated is difficult to measure because there is such a high frequency of both predicted and actual outcomes at repayment equal to 0 and 1. Table 4 displays measures of how

		Predicted		
		Default	**Repaid**	**total**
Observed	**Default**	222	127	348
	Repaid	126	525	652
	total	349	651	

Table 2: Confusion matrix for regression predicting default (decision cutoff = 0.495).

		Predicted		
		Zero	**Nonzero**	**total**
Observed	**Zero**	61	46	107
	Nonzero	65	176	241
	total	126	222	

Table 3: Confusion matrix for regression predicting zero repayment conditional on default (decision cutoff = 0.554).

Percentile	Predicted < Actual	Predicted = Actual	Predicted > Actual
10th	812	186	2
25th	683	304	13
40th	374	391	391
Median	0	652	348

Table 4: Relationship between predicted quantile values and true results for test set. The model does not appear to be empirically well calibrated.

well the model is empirically well calibrated, but it is not immediately apparent what conclusions can be drawn from Table 4. The median predicted value for every loan was repayment = 1, and while the median therefore never underpredicted loan outcome, it was correct for 65.2% of loans. The tenth percentile only overstated the loan outcome for 0.2% of loans, but it correctly predicted loan repayment 18.6% of the time. Taken together, I do not conclude that the posterior predictive distributions of the testing set are empirically well calibrated, but the extent to which they not well calibrated is somewhat ambiguous from the results.

7 Conclusions

To combat increasing default rates and potential future lender discontent, this analysis develops a model to improve the quality of information about loans available to prospective lenders on Zidisha, an international microfinance exchange. The model uses training data to predict posterior distributions of loan outcomes for a testing set of loans, which can then be combined with stated lender risk/reward preferences to determine the lender's optimal interest rate offer. This is done in accordance with the goal to provide borrowers with a low cost of capital in the present while not leaving so many lenders dissatisfied with unprofitable loans that there will be a shortage of capital available for loan in the future. The results, however, indicate that the model may not fit the data very well, which is likely driven by the reality that the that is not very predictive.

Prior to this analysis, lenders made lending decisions and interest rate offers subjectively, based on how they viewed the borrower and the plausibility of his/her proposed idea. This model suggests a transition from fully

subjective loan determination to a decision making process that considers only historical connections between borrower and loan characteristics and loan outcomes. Instead, I believe that the model should be extended to a find middle ground that gives weight to both historical and subjective information. This new model, which is tasked for future research, would be a fully Bayesian model in which prospective lenders specify priors based on their intuition about the borrower. This prior would need to be time efficient and simple enough for borrowers to understand; one possibility might be a two question questionnaire that specifies a lender's beliefs on the projected success of the model and the lender's confidence in their beliefs, each on a scale from 1-10. These two parameters could then be converted into a usable prior. These priors would be combined with historical data in a model similar to the one used in this analysis to yield a Bayesian posterior predictive distribution of loan outcome. Loan outcomes would then be combined with lender risk/reward preferences, as described in this analysis, to determine an optimal interest rate. Given the lack of predictive power in the historical data, the prior specified by prospective lenders might justifiably be weighted relatively heavily in the calculation of the posterior predictive distribution. The resulting Bayesian model would be a valuable addition to Zidisha and potentially other, similar platforms that match borrowers to lenders, insurers to policyholders, or even buyers to sellers when there is uncertainty about the quality of the product that the buyer is purchasing.

References

[1] Bijak, Katarzyna & Thomas, Lyn. *Modelling LGD for unsecured retail loans using sampling methods*. Journal of the Operational Research Society. Published online 12 February 2014.

[2] www.zidisha.org

APPENDIX

The full code used in this analysis is posted to my DropBox on Sakai and otherwise available upon request.