Case 1 Report

Olivia Fan, Mona Su, Shreyas Hallur, Connor Wilchusky

1. Introduction

1.1 Background

Traditional financial institutions have immense financial resources and loan out to enterprises from the international level to the local level, but they often require collateral. Micro-finance is therefore argued by some to be a useful tool to assist with enterprises and people that lack such collateral and still have a need to access capital (Robert Cull & Jonathan Morduch). Given the lack of collateral however, lenders in a micro-finance system take on a risk of default with no possible way for recoupment. It is because of this risk that micro-finance systems needed to minimize the rate of default to remain practical (Kassim, S. H., & Rahman, M.). The purpose of this report is to analyze socio-demographic and geographic factors involved in the risk of default as well as the time to default. Utilizing the knowledge of these factors, the report also seeks to make an estimation about the probability of the default.

The report's analysis will begin with exploratory data analysis to better contextualize the Kiva 2012 database that we are working with as well as a discussion on data processing. From this, to answer our major research questions, we have constructed a variety of models (Exponential AFT, Weibull AFT, Log-Normal AFT) and performed operations with logistic regression and backward selection. Through these operations we have developed results that provide intuition on the interplays between socio-demographic and geographic factors and the risk of/time to default. After presentation of these models we discuss model assumptions and interpretation to provide value to potential investors on the risk of default.

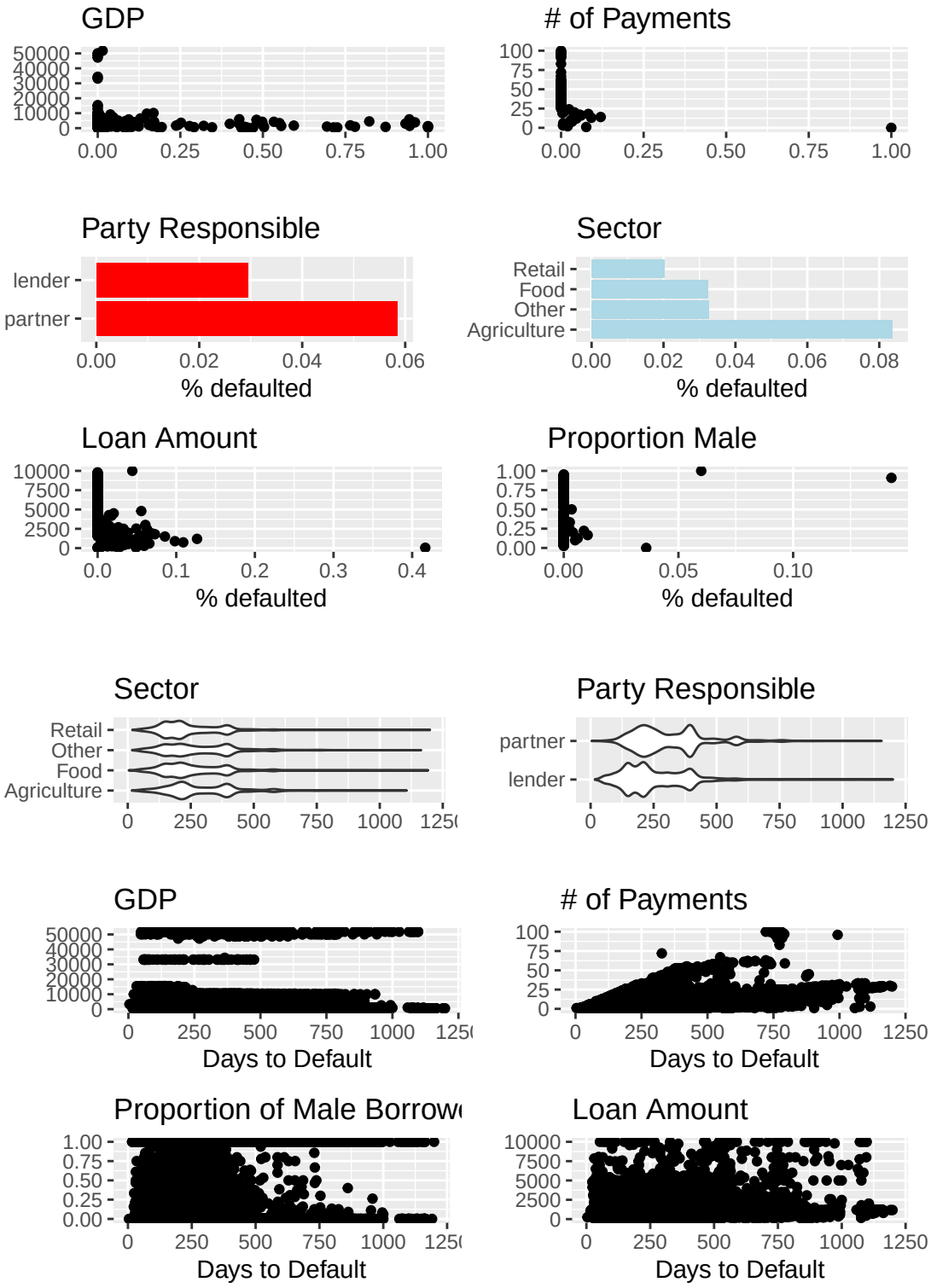
1.2 Definition of Default

For the purposes of this analysis, we defined the time of default as the date at which a borrow fails to pay six scheduled payments in a row. This approach to default closely reflects Kiva's definition of default as a loan that has not received repayments in the last six months. However, the execution of this definition proved somewhat difficult since each borrower can submit a payment of any amount of money at any time, irrespective their planned repayment schedule. To align the borrower's payment and lender's planned repayment schedules, we calculated a borrower's "net balance" at each lender-planned scheduled prepayment date. This net balance variable corresponds to difference between the total amount of settled payments and the sum of the loan repayments due at this time. If the net balance metric remains negative and the borrower has not submitted a settled payment across six scheduled repayment dates, then the loan will meet our rigorous criteria for default.

2. Data Processing

From the four original datasets, we chose to only utilize the loans, loan repayments, loan schedule datasets for analysis, as we reasoned that our clients may not be interested in how their own characteristics inform borrower's likelihood of loan default. In data processing of the loans dataset, we removed the observations with `refunded` status, because these are loans that have been directly refunded and thus never paid back, which does not contribute to investigation of the default status. For all the time and date information, we combined the year, month and day columns into a single date for further analysis and removed the hour, minute and second columns. Then we filtered out extraneous information irrelevant to our analysis, such as the latitude and longitude of the country, the total number of journal entries, Youtube video id, currency exchange loss amount, and the columns showing the language of the comment.

3. Exploratory data analysis



To investigate the effects of macroeconomic factors on loan default, we obtained the countries GDP per capita in the respective years of the loans from the World Bank (2011) API, which reflects the purchasing power of lenders. Furthermore, we clustered the sectors into four larger categories (Agriculture, Food, Retail and

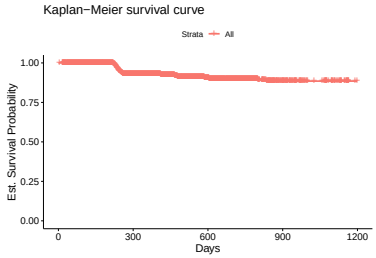
Other) because the former three take up the majority (67%) of the loans. Other variables of interest include the total amount of the loan, GDP per capita of the country where the lender is from, whether partner or lender is liable in case of nonpayment and the number of payments, and the proportion of male lenders in the group. To interpret the effects of interaction terms in the model, we mean centered all the quantitative variables (amount of the loan, GDP, proportion of male lenders and number of payments).

The upper portion of the EDA plots features the correlation of predictors with the default rate, whereas the lower portion features the time to default. We see that loans for which the partner is responsible in case of nonpayment, or loans from the agriculture sector has a significantly higher default rate. Similarly, the violin plots reveal that the liability and sector factors have significant effects on the distribution of time to default. By the same token, we decided to include the predictors from the EDA above in the models.

4. Modeling

To address the the two case study goals separately, we fitted a **survival model** to quantify relationships between sociodemographic/geographic factors and time to default, and a **logistic regression** model to quantify relationships between sociodemographic/geographic factors and risk of default. For each model, we ran the backward selection algorithm to gauge insights into the importance of the predictors, compared the AIC values as well as observed the confidence intervals of the coefficients.

4.1 Survival Analysis (Time to Default)



We decided to fit an accelerated failure time (AFT) model to quantify the relationship between the predictor variables and the time to default. Under an AFT model, the predictor variables we fit will either accelerate or decelerate the time to default. For predictors, we decided to include loan amount, GDP per capita, sector, liability status for nonpayment (lender or partner), and number of scheduled repayments. These covariates were included because we saw from our EDA that they could help explain the variance in default probability. We chose to include an interaction term between loan amount and GDP per capita because we believe people from less wealthier countries might default on larger loans at a faster rate. After fitting the models for the three versions of AFT model (exponential, Weibull and Log-normal), we compared their AIC outputs.

	Exponential	Weibull	Log-normal
AIC	`63078.55`	`67394.25`	`57853.49`

$$log(T_i) = 9.35238 + 0.00018 * loan\_amount - 0.00006 * GDP - 0.71521 * sector\_categoryAgriculture \tag{1}$$

$$+ 0.02725 * sector\_categoryFood + 0.46303 * sector\_categoryRetail \tag{2}$$

$$- 0.88595 * nonpayment\_liabilitypartner + 0.12302 * payments.count + 1.12e^{-08} * loan\_amount * GDP \tag{3}$$

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	6.95581	0.01885	369.00080	0.00000	6.91886	6.99276
loan_amount	0.00005	0.00001	6.16910	0.00000	0.00003	0.00007
GDP	-0.00002	0.00000	-10.15569	0.00000	-0.00002	-0.00001
nonpayment_liabilitypartner	-0.33777	0.01240	-27.22924	0.00000	-0.36208	-0.31346
sector_categoryAgriculture	-0.12897	0.01472	-8.76047	0.00000	-0.15782	-0.10011
sector_categoryFood	-0.02083	0.01489	-1.39908	0.16179	-0.05002	0.00835
sector_categoryRetail	0.08758	0.01685	5.19610	0.00000	0.05454	0.12061
payments.count	0.07095	0.00125	56.94787	0.00000	0.06851	0.07339
loan_amount:GDP	0.00000	0.00000	4.96961	0.00000	0.00000	0.00000
Log(scale)	-0.47368	0.01241	-38.16603	0.00000	NA	NA

Based on the AIC values for the three versions of the ATF model, we decided on the log-normal model with the lowest as the final survival model. All of the parameters and interaction effects in the model have a p-value less than 0.05 and the confidence interval does not include 0, except for the food sector category. All the loans with positive coefficients are expected to survive longer, while those with negative covariates are not expected to survive as long. In this case, survival is defined as not defaulting. For example, holding all else constant, if the loan is for agricultural purposes, it is expected to survive approximately exp(-0.71) or 0.49 times the length of a loan that is not, and the 95% confidence interval spans from a multiplies of 0.45 to 0.53. Additionally, holding all else constant, for every additional payment that has been made, the loan is expected to survive exp(0.12) or 1.13 times longer than the mean length (with a 95% confidence interval of 1.12 to 1.14).

Our model shows that if the partner rather than the investor is responsible for the loan payment, the loan is expected to survive for a shorter amount of time. This could be because if the community partner is responsible, the borrower is more incentivized to pay back the loan. Our model also shows that if a loan is for the agriculture sector, the expected time to default is shorter; while for the retail and food sectors, the expected time to default is longer. This is reasonable since agriculture is a particularly risky field to invest in due to its dependence on external forces (such as weather and pest populations). Next, we can see in our model that with every additional payment that has been made, the loan is expected to survive longer, which is reasonable because the more payments made, the longer the loan has not defaulted.

We thought that it would be reasonable to assume an interaction term between loan amount and GDP because borrowers from countries with lower GDP may default faster if the loan amount is high. Holding all else constant, exp(3.14e-09) is the difference in expected survival times corresponding to an increase in loan amount of 1 dollar for two GDP homogeneous groups which differ by 1 dollar. As for the main predictors, we can say that the expected survival time increases by 1.0001 times corresponding to an increase in loan amount by 1 dollar among countries who have the mean GDP. We can also say that the expected survival time decreases by 0.99 times corresponding to an increase in GDP by 1 dollar among borrowers with the mean loan amount.

4.1.1 Sensitivity analysis on distribution To evaluate the performance of our log-normal AFT model, we performed sensitivity analysis on the choice of distribution. The exponential AFT model output on the data below reveals that all of the predictors in the exponential model have the same direction in coefficients, except for the food sector predictor, which reveals that our AFT model is stable and rarely depends on the choice of the distribution. All of the predictors in our exponential AFT model are still significant in the log-normal model, which reveals that our AFT model is robust to the specifications of the choice of underlying distribution.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	9.35238	0.03817	245.03727	0.00000	9.27758	9.42719
loan_amount	0.00018	0.00002	7.58149	0.00000	0.00013	0.00023
GDP	-0.00006	0.00001	-10.58742	0.00000	-0.00007	-0.00005

term	estimate	std.error	statistic	p.value	conf.low	conf.high
sector_categoryAgriculture	-0.71521	0.04458	-16.04368	0.00000	-0.80259	-0.62784
sector_categoryFood	0.02725	0.05051	0.53950	0.58954	-0.07175	0.12625
sector_categoryRetail	0.46303	0.05906	7.83964	0.00000	0.34727	0.57879
nonpayment_liabilitypartner	-0.88597	0.03723	-23.79879	0.00000	-0.95894	-0.81301
payments.count	0.12302	0.00412	29.85838	0.00000	0.11495	0.13110
loan_amount:GDP	0.00000	0.00000	3.83117	0.00013	0.00000	0.00000

4.2 Logistic Regression (Risk of Default)

Differing from our AFT model, we decided to include the proportion of males in the lender group as another predictor, in light of economics literature that indicates higher default rates associated with male lenders such as found by researchers at Barcelona Graduate School of Economics (Moltalvo & Reynal-Querol, 2020).

All of the predictors are statistically significant and the confidence intervals do not include zero, except for the sector category of food. Here, a negative coefficient indicates that the predictor decreases the odds of the loan defaulting, while a positive coefficient indicates the opposite. For example, the loan amount coefficient can be interpreted as holding all else constant, for every increase in dollar amount of the loan, the odds of a default decreases by exp (-0.00028) or 0.99 times among borrowers from countries with an average GDP and average number of payments. This could be because borrowers who take on higher loans are more confident that they are capable of paying it back. Additionally, we can conclude from the model that holding all else constant, with every additional increase in the percentage of male borrowers, the odds of a default increases by exp(0.45) or 1.56 times. This could be because male borrowers are less trustworthy than female borrowers.

We can see that with every dollar increase in GDP of a country that the loan is based in, the odds of defaulting increases by 1.0006 times among people who have the average loan amount. This could be because there is a higher chance of risky loans in richer countries. Additionally, for loans based in either the agriculture or food sector, the odds of defaulting increases, while the opposite is true for loans based in retail. This could be because the former two are riskier fields to make investments in than the latter due to its uncertainty. Lastly, for every additional payment that the borrower has made, the odds of defaulting decreases for borrowers who have the average loan amount.

$$log(\frac{P}{1-P}) = -4.02712 - 0.00028 * loan\_amount + 0.45698 * proportion\_males - 0.00006 * GDP \quad (4)$$

$$+ 0.95310 * sector\_categoryAgriculture + 0.00751 * sector\_categoryFood \quad (5)$$

$$- 0.47920 * sector\_categoryRetail + 1.03806 * nonpayment\_liabilitypartner \quad (6)$$

$$- 0.07067 * payments.count - 1.105e^{-08} * loan\_amount \cdot GDP \quad (7)$$

$$- 0.00003 * loan\_amount \cdot payments.count \quad (8)$$

We then elected to evaluate two interaction terms between loan amount and GDP per capita and between loan amount and the number of payments. People from less wealthy nations may have higher default rates for larger loans while people who budget shorter schedules for larger loans may be more likely to default. The first interaction between loan amount and GDP is the difference between the log-odds ratios corresponding to an increase in loan amount of 1 dollar for two GDP homogeneous groups which differ by 1 dollar. It can also be interpreted as the difference between the log-odds ratios corresponding to an increase in GDP of 1 dollar for two loan amount homogeneous groups which differ by 1 dollar. The second interaction between loan amount and number of payments can be interpreted similarly as the difference between the log-odds ratios corresponding to an increase in loan amount of 1 dollar for two number-of-payments homogeneous groups which differ by one payment.

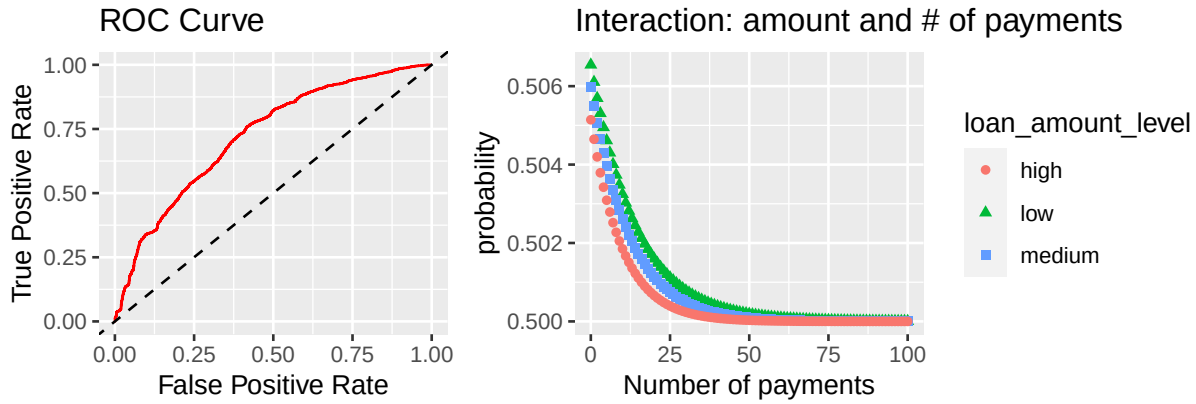
We then ran a F-test to assess the inclusion of interactions terms and decided to include both in the full model. After adding the interaction terms, we performed backward selection and yielded the same full model, therefore all the predictors are significant.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	-4.02712	0.04375	-92.05571	0.00000	-4.11355	-3.94205
loan_amount	-0.00028	0.00003	-9.97942	0.00000	-0.00034	-0.00023
proportion_males	0.45698	0.03923	11.64936	0.00000	0.37991	0.53369
GDP	0.00006	0.00001	10.36988	0.00000	0.00005	0.00007
sector_categoryAgriculture	0.95310	0.04574	20.83958	0.00000	0.86364	1.04294
sector_categoryFood	0.00751	0.05204	0.14440	0.88519	-0.09467	0.10934
sector_categoryRetail	-0.47920	0.06036	-7.93887	0.00000	-0.59830	-0.36163
nonpayment_liabilitypartner	1.03806	0.03939	26.35496	0.00000	0.96090	1.11530
payments.count	-0.07067	0.00430	-16.44257	0.00000	-0.07916	-0.06232
loan_amount:GDP	0.00000	0.00000	-3.61618	0.00030	0.00000	0.00000
loan_amount:payments.count	-0.00003	0.00001	-5.79631	0.00000	-0.00004	-0.00002

Resid. Df	Resid. Dev	Df	Deviance	F	Pr(>F)
85895	26101.36	NA	NA	NA	NA
85894	26080.74	1	20.61592	20.61592	5.6e-06
85893	26045.42	1	35.32630	35.32630	0.0e+00

The model yields an AUC (Area Under the Curve) on the ROC plot of 0.7238, which indicates that there is a 72.38% probability that the classifier will rank a randomly chosen default example higher than a randomly chosen non-default example, indicating a moderate to good performance in binary classification.

4.2.1 Interaction Term Interpretation



We found a significant interaction term between the number of payments and the loan amount, with the term having small p values, and confidence interval significantly different from 0 in both the survival and logistic models. We visualized the effect of number of payments on the probability of defaulting under different levels of `loan_amount` values. Under high value of loan amount (1000 dollars), the probability of default decreases more drastically with the number of payments. Whereas, under low value of loan amount (25 dollars), the number of payments has a smaller effect on this probability. Under medium value of loan amount (400 dollars), the number of payments has a moderate effect on decrease in the probability of default as number of payments increases.

4.2.2 K-Fold Stratefied Cross Validation To apply cross validation to the data set with imbalanced class distribution, we performed 5-fold stratified cross validation. We obtained a mean accuracy of 96.14% which demonstrates that the overall estimated accuracy of the model is very high. We obtained a standard deviation of accuracy of 2.941729e-05, which demonstrates that the variability of the accuracy across all folds is low, thus indicates that the accuracy estimate is more stable.

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
Accuracy	0.9614086	0.9613549	0.9613549	0.9614086	0.9614086

4.3 Model Assumptions

4.3.1 Survival Analysis Assumptions

Firstly, it is reasonable to assume that the survival probabilities are the same for all the samples who joined late in the study and those who have joined early since the probability to default is not dependent on when the loan is recorded in the study. Second, the assumption that the occurrence of event is done at a specified time is met because using our definition of default, the event occurs on a specific day. Thirdly, the assumption that censoring is independent or unrelated to the likelihood of developing the event of interest because our definition of default is only dependent on if borrower misses payments.

4.3.2 Logistic Regression

Randomness: It is reasonable to assume the observations in the dataset are randomly selected and are representative of the population of interest.

Independence: It is reasonable to assume that the observations are independent of one another.

Linearity: We can see that all of the residuals for the predictor variables used in the logistic regression do not have a pattern in their binned residual distributions. The residual distribution for the proportion of male borrowers appear a bit abnormal, but the outlines at 0 and at 1 makes sense because there are a significant number of loans that are only male or only female borrowers. We can see for the categorical variables, all the residuals are very close to zero; therefore, the linearity assumption is satisfied.

5. Sensitivity Analysis

Due to the rigorousness and flexibility of our calculation of time to default, we were able to perform sensitivity analysis by varying our definition of default.

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	7.58062	0.04277	177.23702	0.00000	7.49679	7.66445
loan_amount	0.00019	0.00002	10.53533	0.00000	0.00016	0.00023
GDP	-0.00006	0.00001	-10.99320	0.00000	-0.00007	-0.00005
sector_categoryAgriculture	-0.67687	0.03177	-21.30605	0.00000	-0.73914	-0.61461
sector_categoryFood	-0.03366	0.03468	-0.97032	0.33189	-0.10164	0.03433
sector_categoryRetail	0.28822	0.03891	7.40815	0.00000	0.21197	0.36448
nonpayment_liabilitypartner	-1.26724	0.02586	-48.99852	0.00000	-1.31794	-1.21655
payments.count	0.14416	0.00296	48.62945	0.00000	0.13835	0.14997

term	estimate	std.error	statistic	p.value	conf.low	conf.high
loan_amount:GDP	0.00000	0.00000	3.53627	0.00041	0.00000	0.00000

Based on the model output derived from the alternative definition of default as a loan that has not received repayments in the last three months instead of six months, we refitted the model. We observe that the directional impact of all statistical significant main effect and interaction terms remains the same for both definitions of default. For instance, both definitions of default suggest that lending for agricultural purposes accelerates the time to default. Therefore, our model is robust to specifications of the definition of default.

6. Implications & Model Interpretation

The directions of all the main effect terms correspond with one of another in both models. If the coefficient is positive in the AFT model, it is negative in the logistic, which means that for that given predictor, we can conclude the loan is expected to survive longer before defaulting from the AFT model and has a lower chance of default from the logistic model. The predictors associated with a longer time to default and a lower risk of default involve borrowers from the retail sector, larger loans, country of origin with greater GDP per capita and loans with a greater number of payments. Both results meet our expectations, as borrowers from the retail sector may have greater financial stability given their diverse product offerings. Likewise, larger loans may provide borrowers more capital to generate revenue for repayments while loans divided into a greater number of scheduled payments may prove more manageable for borrowers. The predictors associated with a shorter time to default and greater risk of default are smaller loan amounts, borrowers in the agriculture sector, liability held by partner rather than an investor, and country of origin with lower GDP per capita. Increases in the interaction term between loan amount and GDP per capita are further associated with a shorter time to default and a higher risk of default. In more model-specific results, a greater proportion of male borrowers is associated with greater risk of default in the logistic regression model. However, increases in the interaction term between loan amount and the number of payments are associated with a lower risk of default in the same model.

7. Limitations & Future Work

Our statistical analysis protocol yields a number of important benefits. Firstly, our two-pronged approach involving both survival analysis and logistic regression allows us to separately identify predictors of time to default and risk of default, respectively, all while preserving interpretability. Secondly, our analysis plan benefits from an independent definition of default that draws heavily from the official Kiva model. Post-hoc sensitivity analyses have further indicated that our definition of default is relatively robust and yields the same results regardless of cutoff window. A final strength stems from our holistic approach to model selection, wherein AIC comparisons and backward selection have enabled us to select the most effective model.

One limitation is that our code did not include losses from currency exchange, which could complicate a small number of loan repayment calculations. Another minor setback stems from our code's computation time. Since we had to compute balances for each scheduled payment dates, irrespective of the amount, frequency, or timing of the borrower's repayments, the temporal comparisons across the schedule and repayment files added some added inefficiencies. However, the storage of computationally difficult results in separate csv files has reduced our run time to 4:35. A final limitation stems from our reliance on settled payments in our definition of default which depends on the processing time across countries.

Future analyses may want to consider lender behavior and see if certain patterns may influence a borrower's default rates. Another room for improvement stems from the reality that most borrowers finish paying off their loans early and thus naturally eliminate their risk for default. We have already computed time to payment in an attempt to account for this possibility, but survival analysis models that allow for the possibility of early payment, such as mixture cure models, proved too difficult to implement given their lack of documentation and prior financial applications.

8. Citations

1.

Robert Cull & Jonathan Morduch. (2017). *Microfinance and Economic Development* (Policy Research Working Paper 8252). https://documents1.worldbank.org/curated/en/107171511360386561/pdf/WPS8252.pdf

2.

Kassim, S. H., & Rahman, M. (2018). Handling default risks in microfinance: The case of bangladesh. *Qualitative Research in Financial Markets, 10*(4), 363-380. doi:https://doi.org/10.1108/QRFM-03-2017-0018

3.

Jose G. Moltalvo & Marta Reynal-Querol. (2020). *Gender and Credit Risk: A View From the Loan Officer's Desk*. Barcelona Graduate School of Economics. https://bse.eu/sites/default/files/working_paper_pdfs/1076_0.pdf

4.

The World Bank. (2011). *The World Bank Data Bank*. https://databank.worldbank.org/databases/page/1?qterm=GDP